Estimating Average Treatment Effects Using a Modified Synthetic Control Method: Theory and Applications

Kathleen T. Li*
Marketing Department
The Wharton School, the University of Pennsylvania

June 20, 2017

Abstract

The synthetic control method, a powerful approach for estimating average treatment effects (ATE), is an important evaluation tool that is widely used in statistics, economics, marketing and other social sciences. In this paper we seek to estimate ATE based on the synthetic control method using panel data with large pre and post-treatment observations. The numbers of treated and control units are fixed, though. Up to now, there has been no formal inference theory in this situation. Thus, our main contribution is deriving the asymptotic distribution of the synthetic control ATE estimator. The asymptotic distribution is non-normal, non-standard, and the standard bootstrap does not work, but we show that a carefully designed sub-sampling method can be applied to obtain valid inferences. We also show that simple modifications proposed by Doudchenko and Imbens (2016) make the synthetic control method applicable to a wider range of data generating processes. Simulations and an empirical application demonstrate the usefulness of the modified method when treatment and control units are drawn from heterogenous distributions.

*Katheen T. Li is a doctoral candidate in marketing at the Wharton School, the University of Pennsylvania. She acknowledges co-advisers David R. Bell and Christophe Van den Bulte for invaluable guidance and thanks committee members Eric T. Bradlow and Dylan S. Small for helpful comments. In addition, she is grateful for David R. Bell and Warby Parker for providing the data. Correspondence regarding this manuscript can be addressed to Kathleen T. Li, katli@wharton.upenn.edu, The Wharton School, 700 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104.

1 Introduction

Identifying average treatment effects (ATE) from non-experimental data has become one of the most important endeavors of social scientists over the last two decades. It has proven to be one of the most challenging as well. The difficulty lies in accurately estimating the counterfactual outcomes for the treated units in the absence of treatment. Early literature on examining treatment effects focused on evaluating the effectiveness of education and labor market programs (Ashenfelter 1978, Ashenfelter and Card 1985). More recently, marketing and management scientists have used quasi-experimental data to evaluate such treatment effects as Internet information on financing terms for new cars (Busse, Silva-Russo and Zettlemeyer 2006); price reactions to rivals local channel exits (Ozturk, Venkataraman and Chintagunta, 2016), and offline bookstore openings on sales at Amazon (Forman, Ghose and Goldfarb 2009). Others have used Difference-in-Differences (DID) methods to examine various treatment effects, especially in digital environments, such as how online reviews drive sales (Chevalier and Mayzlin 2006); executional strategies for display advertising (Goldfarb and Tucker 2011); online information on consumers' strategic behavior (Mantin and Rubin 2016); and how offline stores drive online sales (Wang and Goldfarb 2017).

DID and the propensity score matching methodologies are perhaps the most popular approaches used to estimate treatment effects. These methods are especially effective when there are large number of treatment and control units over short time periods. One crucial assumption for the DID method is that outcomes of the treated and control units follow parallel paths in the absence of treatment. Violation of this parallel lines assumption in general will result in biased DID estimates. For panel data with a relatively large number of time-series observations, alternative methods may be better suited than DID for estimating counterfactual outcomes. For example, the synthetic control method proposed by the seminal work of Abadie and Gardeazabal (2003), and Abadie, Diamond and Hainmeller (2010) can be used successfully to estimate average treatment effects (ATE). This method has many attractive features: First, it is more general than the conventional difference-in-differences method, because it allows for different control units to have different weights (individual specific coefficients) when estimating the counterfactual outcome of the treated unit. Second, the synthetic control method restricts the weights assigned to the control group to be non-negative (because outcome variables are likely to be positively correlated), it may lead to better extrapolation. Finally, the synthetic control method can be adjusted to work even when the number of time periods is small. Athey and Imbens (2016) described the synthetic control method as arguably the most important innovation in the evaluation literature in the last 15 years.

Abadie, Diamond and Hainmueller have suggested that the potential applicability of the synthetic control method to comparative case studies is very broad (Abadie, Diamond and Hainmueller 2010, page 493). However, this method is not without some limitations. For example, the restriction that the sum of the weights assigned to the control equal to one implicitly requires that outcomes for the

treated unit and a weighted average of control units have followed parallel paths over time in the absence of treatment (e.g., assumption 3.1 in Abadie (2005)). When panel data contains a long time series, the condition that the weights sum up to one can be restrictive and lead to poor fit. This occurs when the 'parallel lines' assumption is violated. For many social science datasets, the 'parallel lines' assumption may not hold. Note that the requirement for outcomes of control and treatment units to follow parallel paths is weaker than requiring random assignment (i.e. random assignment means that control and treatment units are randomly selected from the same population, implying that they follow parallel paths).

In this paper we consider both the standard synthetic control method and a modified synthetic control method suggested by Doudchenko and Imbens (2016). Their proposed modifications are: adding an intercept, and dropping the slope coefficients sum-to-one restriction in a standard synthetic control model. Dropping the latter restriction makes the modified method applicable to a wider range of data settings. That is because when the sample paths of treatment and the weighted sum of the control units (with weights sum to one) are not parallel in the pre-treatment period, the standard synthetic control method leads to poor in-sample fit. One should not use the method in such a case (e.g., Abadie, Diamond and Hainmueller 2010, page 495). However, we show that even in this case, it is likely that the modified synthetic control can be used to deliver reliable ATE estimation results. That is because the modified synthetic control method can adjust the slope of a linear combination of the control outcomes and make it parallel to the treated unit's outcome sample path (in the absence of treatment). We use simulated and real data to demonstrate the improvement of the modified synthetic control over the standard synthetic method when the treated and control units are drawn from heterogeneous distributions.

Most existing work based on the synthetic control ATE estimator relies on the assumption that the treatment units are random assigned and uses placebo tests to conduct inferences. Hahn and Shi (2016) show that the validity of using placebo tests requires a strong normality distribution assumption for the idiosyncratic error terms under a factor model data generating framework. Conley and Taber (2011), and Ferman and Pinto (2016a, b) propose rigorous inference methods for DID and synthetic control ATE estimators under different conditions. Conley and Taber (2011) assume that there is only one treated unit and a large number of control units, and that the idiosyncratic errors from the treated and the control units are identically distributed (a sufficient condition for this is the random assignment to the treated unit). They show that one can conduct proper inference for the DID ATE estimator by using the control units' information. Their method allows for both the pre and the post-treatment periods to be small. Assuming instead that the pre-treatment period is large and the post-treatment period is small, Ferman and Pinto (2016a, b) show that Andrews' (2003) end-of-sample instability test can be used to conduct inference for ATE estimators without requiring the random assignment to the treated unit assumption. However, when both pre-treatment time period,

T_1, and post-treatment period, T_2, are large, Andrews' test becomes invalid because an estimation error of order $\sqrt{T_2/T_1}$ becomes non-negligible. Apparently, without imposing the random assignment assumption (for selecting the treated units), there is no rigorous inference theory for the synthetic control ATE estimator if both the numbers of treated and control units are fixed and finite, but both the pre and post-treatment periods are large. This paper fills that gap. Using the projection theory (e.g., Zarantonello (1971), Fang and Santos (2015)), we derive the asymptotic distributions of the standard and the modified synthetic control ATE estimators under the assumption that both T_1 and T_2 are large. The asymptotic distribution is non-normal and non-standard. Moreover, it is known that the standard bootstrap does not work for the synthetic control estimator (Andrews (2000), Fang and Santos (2015)). Fortunately, we are able to show that a carefully designed sub-sampling method provides valid inferences. We also apply our new theoretical results to conducting inferences for empirical data. We estimate the effect of WarbyParker.com's showroom (at Columbus, Ohio) opening on its sales. For this data, $T_1 = 90$ and $T_2 = 20$. Using simulations for this T_1, T_2 combination, the inference based on our proposed sub-sampling method yields more accurate estimated confidence intervals than the estimates using Andrews' instability test. The reason is that $T_2 = 20$ is not negligible compared to $T_1 = 90$, rendering Andrews' test improper for our empirical data.

In short, we make three contributions in this paper. First, we show via simulations and an empirical example that a modified synthetic control method, which is robust to 'non-parallel paths' situations, greatly enhances the applicability of the synthetic control method to estimating ATE. Second, under the assumption that there are large pre and post-treatment observations, we derive the asymptotic distribution of the modified synthetic-control-method ATE estimator. The asymptotic distribution is non-normal and non-standard. Third, we propose an easy-to-implement sub-sampling method and show that it leads to valid inferences. In addition, we provide a simple sufficient condition under which the synthetic control estimator has a unique global solution.

The remaining parts of the paper are organized as follows. In section 2, we discuss two existing methods for estimating ATE: the DID method and the standard synthetic control method. In section 2.3, we consider the modified synthetic control ATE estimator as suggested by Doudchenko and Imbens (2016), and discuss a condition for the uniqueness of the estimator. In Section 3 we derive the asymptotic distribution of the synthetic-control-method-based average treatment effects estimator, while in Section 4 we propose using a simple sub-sampling method to conduct inference. Section 5 reports simulation results examining the effectiveness of using the sub-sampling method in inferences. Section 6 presents an empirical application that examines the average treatment effects of opening a physical showroom on WarbyParker.com's sales. Finally, Section 7 concludes the paper.

2 Estimating ATE using panel data

We start by introducing some notation and discussing two methods of estimating average treatment effects using panel data. We first discuss the popular Difference-in-Differences (DID) method and then the synthetic control method of Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010).

Let y_{it}^1 and y_{it}^0 denote unit i's outcome in period t with and without treatment, respectively. The treatment effect from intervention for the i^{th} unit at time t is defined as

$$\Delta_{it} = y_{it}^1 - y_{it}^0. \tag{2.1}$$

However, we do not simultaneously observe y_{it}^0 and y_{it}^1. The observed data is in the form

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0, \tag{2.2}$$

where $d_{it} = 1$ if the i^{th} unit is under the treatment at time t, and $d_{it} = 0$ otherwise.

We consider the case that there is a finite number of treated and control units and the treated units are drawn from heterogenous distributions (i.e., they not randomly assigned). Also, the treatment time occurs at different times to different treated units. In this type of situations, it is reasonable to estimate ATE (over post-treatment period) for each treated unit separately. In this way, one can obtain ATE for each treated unit. If one also wants to obtain ATE over all the treated units, one can average (possibly with different weights) over all individual of the treated group. In contrast, if one averages over treated individuals at the beginning of the estimation stage, then individual ATE information cannot be recovered and the valuable heterogeneous individual ATE information will be lost. For these reasons, in this paper we focus on the case that there is one treated unit that receives a treatment at time $T_1 + 1$. Without loss of generality we assume that this is the first unit so that we have $y_{1t} = y_{1t}^1$ for $t \geq T_1 + 1$. The difficulty in estimating the treatment effects $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$ is that y_{1t}^0 is not observable for $t \geq T_1 + 1$. Assuming that y_{1t} is correlated with y_{jt} for $j > 1$ and that there is no treatment applied to any other units (except the first unit) for all $t = 1, ..., T_1, T_1 + 1, ..., T$, we can construct an estimator for the unobserved y_{1t}^0. Specific methods for estimating y_{1t}^0 are discussed in subsequent sections. For now, let $\hat{y}_{1t}^0$ be a generic estimator of y_{1t}^0. Then the treatment effect at time t is estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ $(t = T_1 + 1, ..., T)$ and the average treatment effect, averaging over the post-treatment period, is estimated by (where $T_2 = T - T_1$)

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t}. \tag{2.3}$$

2.1 The Difference-in-Differences method

In this section, we discuss the Difference-in-Differences estimation method. Using the same notation as in the previous section, y_{1t} denotes the outcome of unit 1 at time t. The difference of average outcomes after and before the treatment date is given by

$$ATE_{1,DID} = \frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{1t} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{1t}, \tag{2.4}$$

where T_1 is the pre-treatment sample size, and $T_2 = T - T_1$ is the post-treatment sample size. Similarly, the difference of outcomes for the $N-1$ control units after and before the treatment intervention date is computed by

$$ATE_{control,DID} = \frac{1}{N-1}\sum_{j=2}^{N}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{jt} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{jt}\right]. \tag{2.5}$$

The DID average treatment effect is calculated via Difference-in-Differences method:

$$\begin{aligned} ATE_{1,DID} &= ATE_{1,DID} - ATE_{control,DID} \\ &= \frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{1t} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{1t} - \frac{1}{N-1}\sum_{j=2}^{N}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{jt} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{jt}\right]. \end{aligned} \tag{2.6}$$

The intuition behind the DID method is that y_{jt}, $j = 1, ..., N$, are random draws from a homogenous population. Therefore, $\bar{y}_{control,t} = \sum_{j=2}^{N} y_{jt}/(N-1)$ may mimic y_{1t} well in the absence of treatment. In order to improve the fit of using $\bar{y}_{control,t}$ to approximate y_{1t}, we add an intercept term δ_1 to $\bar{y}_{control,t}$ and use $\delta_1 + \bar{y}_{control,t}$ to approximate y_{1t} for $t \leq T_1$. Naturally, we estimate δ_1 using the pre-treatment data by

$$\hat{\delta}_1 = \bar{y}_1 - \bar{y}_{control} = \frac{1}{T_1}\sum_{t=1}^{T_1} y_{1t} - \frac{1}{T_1}\sum_{t=1}^{T_1}\frac{1}{N-1}\sum_{j=2}^{N} y_{jt}, \tag{2.7}$$

where $\hat{\delta}_1$ is the least squares estimator of δ_1 in $y_{1t} - \bar{y}_{control,t} = \delta_1 + error_t$. Therefore, the DID fitted value is given by

$$\hat{y}^0_{DID,1t} = \hat{\delta}_1 + \frac{1}{N-1}\sum_{j=2}^{N} y_{jt}, \qquad t = 1, ..., T_1, T_1+1, ..., T \tag{2.8}$$

where $\hat{\delta}_1$ is given in (2.7). Note that $\hat{y}^0_{DID,1t}$ gives the in-sample fitted value for $t \leq T_1$; and it gives out-of-sample counterfactual estimated curve for $t \geq T_1 + 1$.

When there is only one (or only a few) unit that receives a treatment, and there are many control units, the standard inference theory may fail and some modifications are needed. See Ferman an Pinto (2016a) for specific modifications that are needed to give valid inference for the DID estimator defined in (2.6).

2.2 The Synthetic Control method

We continue to examine the scenario where a treatment was administered to the first unit at $t = T_1+1$. Thus, the remaining $N-1$ units are control units. In order to use a unified notation to cover both the synthetic control and the modified synthetic control methods, we add an intercept to the standard synthetic control method. Therefore, utilizing the correlation between y_{1t} and y_{jt} (say, all outcomes are correlated with some common factors), $j = 2, ..., N$, one can estimate the synthetic control counterfactual outcome y_{1t}^0 based on the following regression model.

$$y_{1t} = x_t'\beta_0 + u_{1t}, \qquad t = 1, ..., T_1, \tag{2.9}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$ is an $N\times 1$ vector of the control units' outcome variables, $\beta_0 = (\beta_{0,1}, ..., \beta_{0,N})$ is an $N \times 1$ vector of unknown coefficients, and u_{1t} is a zero mean, finite variance idiosyncratic error term.

Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010) propose a synthetic control method that uses a weighted average of the control units to approximate the sample path of the control unit. The weights are selected by best fitting the outcome of the treated unit using pre-treatment data, and the weights are non-negative and sum to one. Specifically, one selects $\beta = (\beta_1, ..., \beta_N)'$ via the following constrained minimization problem:

$$\hat{\beta}_{T_1,Syn} = \arg \min_{\beta\in\Lambda_{Syn}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t'\beta\right]^2, \tag{2.10}$$

where

$$\Lambda_{Syn} = \{\beta \in \mathcal{R}^N : \ \beta_j \geq 0 \text{ for } j = 2, ..., N \text{ and } \textstyle\sum_{j=2}^N \beta_j = 1\}. \tag{2.11}$$

With $\hat{\beta}_{T_1,Syn}$ defined as the minimizer to (2.10), the synthetic control fitted/predicted curve is

$$\hat{y}_{1t,Syn}^0 = x_t'\hat{\beta}_{T_1,Syn}, \qquad t = 1, ..., T_1, T_1+1, ..., T. \tag{2.12}$$

Note that $\hat{y}_{1t,Syn}^0$ is the in-sample fitted curve for $t = 1, ..., T_1$, and $\hat{y}_{1t,Syn}^0$ gives the predicted counterfactual outcome of y_{1t}^0 for $t = T_1 + 1, ..., T$. The ATE is estimated by

$$\hat{\Delta}_{1,Syn} = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t,Syn}^0).$$

It can be shown that, when the number of treated units is larger than the pre-treatment period, an unique weight vector β that minimizes (2.10) may not exist. In such cases, it is necessary to regulate the weights such as imposing non-negativity and sum to one restrictions. The rationale of imposing non-negativity restriction is that in most applications, y_{jt}'s are positively correlated with each other,

and therefore they tend to move up or down together. The add-to-one restriction $\sum_{j=2}^{N} \beta_j = 1$ introduced by Abadie, Diamond and Hainmueller (2010) implicitly assumes that a weighted average outcomes for the control units and the treated unit's outcome would have followed parallel paths over time in the absence of treatment. The restriction that the slope coefficients sum to one can improve the out-of-sample extrapolation when the "parallel lines" assumption holds. When X, the $T_1 \times N$ control units' data matrix, has a full column rank (this requires that the pre-treatment period T_1 is larger than the number of control units), we show in Appendix A that the constrained weight vector β, as a minimizer to (2.10), is unique. In this case, the zero intercept and the slope coefficient sum to one restrictions should be considered on their merit rather than a rule as discussed in Doudchenko and Imbens (2016).

Since our main interest is to forecast y_{1t}^0 for $t \geq T_1 + 1$ rather than in-sample-fit, as long as T_1 is moderately large, we recommend using $N < T_1$ control units in estimating $\hat{y}_{1t}^0$. There are at least two reasons for doing this: (i) When treated and control outcomes are generated by a fixed number of common factors, it can be shown that using a finite number of control units give more accurate predicted counterfactual outcome than using a large number of control units. This reason is quite intuitive as it is well known that using too many regressors in a forecasting model leads to large prediction variance; (ii) when $N > T_1$, $\hat{\beta}_{T_1,Syn}$ cannot be uniquely determined in general. In practice when one faces a large number of control units, one can use AIC, BIC, LASSO (Bradley, Hastie, Johnstone and Tibshirani 2004), or the best subset selection method proposed by Doudchenko and Imbens (2016) to select significant control units.

Abadie, Diamond and Hainmueller (2010) also suggest using covariates to improve the fit when relevant covariates are available. Adding covariates to the model is straightforward. To focus on the main issue of the paper, we will first consider the case without any relevant covariates and discuss how to add relevant covariances in the empirical application section 6. When the treatment unit's outcome and a weighted average of the control units' outcome do not follow parallel pathes in the absence of treatment, the standard synthetic control method may lead to biased estimation results. In section 6 we show that a real data example exhibits this behavior. If one mis-applies the standard synthetic control method to such a data case, one would obtain a severely biased estimation results. This belongs to a scenario that Abadie, Diamond and Hainmueller (2010, page 495) cautioned that one should not apply the standard synthetic control method. However, we will show in Section 6 that the modified estimation method can substantially reduce the estimation bias, giving a reliable ATE estimation result for this empirical data.

2.3 The Modified Synthetic Control method

For many non-experimental data used in economics, marketing and other social science fields, the treated unit and the control units may exhibit substantial heterogeneity and the treated unit's outcome

and a weighted average (with weights sum to one) of the control units' outcomes may not follow parallel paths in the absent of treatment. In this section, we consider simple modifications as advocated by Doudchenko and Imbens (2016). Specifically, we add an intercept and remove the coefficients sum to one restriction in a standard synthetic control model, i.e., we still keep the non-negative constraints: $\beta_j \geq 0$ for $j = 2, ..., N$, but drop the restriction $\sum_{j=2}^{N} \beta_j = 1$. When the sum of the estimated weights (coefficients) is far from one, we suggest not to impose the add-to-one restriction. Therefore, Doudchenko and Imbens' modified synthetic control method is the same as (2.10) except that the add-to-one restriction on the slope coefficients is removed, i.e., one solves the following (constrained) minimization problem:

$$\hat{\beta}_{T_1,Msyn} = \arg \min_{\beta \in \Lambda_{Msyn}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t' \beta \right]^2, \tag{2.13}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$, β is an $N \times 1$ vector of parameters, and

$$\Lambda_{Msyn} = \{\beta \in \mathcal{R}^N : \ \beta_j \geq 0 \text{ for } j = 2, ..., N\}. \tag{2.14}$$

Let X be the $T_1 \times N$ matrix with its i^{th} row given by $x_t' = (1, y_{2t}, ..., y_{Nt})$, we show in Appendix A that when X has full column rank (which requires that $T_1 \geq N$), the (modified) synthetic control minimizers $\hat{\beta}_{T_1,Syn}$ and $\hat{\beta}_{T_1,Msyn}$ are uniquely defined.

With $\hat{\beta}_{T_1,Msyn}$ defined above, the counterfactual outcome is estimated by $\hat{y}_{1t}^0 = x_t' \hat{\beta}_{T_1,Msyn}$ for $t = T_1 + 1, ..., T$, and the ATE is estimated by

$$\hat{\Delta}_{1,Msyn} = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left[y_{1t} - \hat{y}_{1t}^0 \right]. \tag{2.15}$$

3 Distribution Theory for the Synthetic Control ATE Estimator

3.1 Placebo Tests

Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010) offer no formal inference theory. Instead, they conduct placebo tests. Placebo tests evaluate the significance of estimates by answering the question: under the null hypothesis of no treatment effect, how often would we obtain an ATE estimate of a certain large magnitude purely by chance? Essentially, a placebo test involves demonstrating that the effect does not exist when it is not expected to exist. In the case of the synthetic control method, we first apply the synthetic control method to the treatment unit and calculate the treatment effect. Then, we pretend that one of the control units (that did not receive treatment) is the treatment unit and apply the synthetic control method. In this case, we expect the "treatment effect" to be close to zero. This can be done iteratively for all the control units to obtain the "treatment effect". By plotting the treatment effects for the treatment unit and control units together, we should see that the treatment effect for the true treatment unit has the greatest magnitude.

Abadie and Gardeazabal (2003) provide placebo test plots for California (treatment unit) and 38 control units. However, there were several control units whose treatment effects were of greater magnitude than California. In order to address this concern, Abadie and Gardeazabal (2003) argue that for certain control units, synthetic control method does not fit well enough in-sample (as measured by the pre-intervention mean squared prediction error, MSPE). Therefore, they show three additional placebo tests plots discarding states with MSPE twenty times higher, five times higher, and two times higher than California, respectively. As expected, the placebo test plots look progressively better. However, in practice, it is difficult to decide what the MSPE cutoff should be. In addition, placebo tests can only provide correct confidence intervals under quite stringent conditions (Hahn and Shi 2016). To date, the majority of papers that use synthetic control also use placebo tests.

We provide the formal inference theory and show that subsampling can be used to obtain confidence bounds and conduct hypothesis tests. To develop the inference theory, we use the theory of projections onto convex sets. In the next few sections, we first formally present the synthetic control method and modified synthetic control method as an optimization problem over a convex set and derive the asymptotic distributions in order to facilitate formal inference.

3.2 A projection of the unconstrained estimator

To study the distribution theory of the synthetic control ATE estimator, we will first show that one can express the constrained estimator as a projection of the unconstrained (the ordinary least squares) estimator onto a constrained set. Then we use the theory of projection onto convex sets to derive the asymptotic distribution of the synthetic control ATE estimator.

Let $\hat{\beta}_{OLS}$ denote the ordinary least squares estimator of β_0 using data $\{y_{1t}, x_t\}_{t=1}^{T_1}$. We show in Appendix A that the constrained estimator $\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda} \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ can be obtained as a projection of $\hat{\beta}_{OLS}$ onto the convex set Λ, where $\Lambda = \Lambda_{Syn}$ or $\Lambda = \Lambda_{Msyn}$.

We first define some projections. For $\theta \in \mathcal{R}^N$, we define two versions of projection of θ onto a convex set Λ as follows:

$$\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}(\theta-\lambda)'(X'X/T_1)(\theta-\lambda), \tag{3.1}$$

$$\Pi_{\Lambda}\theta = \arg\min_{\lambda\in\Lambda}(\theta-\lambda)'E(x_tx_t')(\theta-\lambda). \tag{3.2}$$

Here we use the notation Π_Λ to denote a projection onto the set Λ. Note that the first projection Π_{Λ,T_1} is with respect to a random norm $\|a\|_X = \sqrt{a'(X'X/T_1)a}$, while the second projection Π_Λ is with respect to a non-random norm $\|a\|_E = \sqrt{a'E(x_tx_t')a}$, i.e., $\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}||\lambda-\theta||_X^2$ and $\Pi_\Lambda\theta = \arg\min_{\lambda\in\Lambda}||\lambda-\theta||_E^2$. The first projection will be used to connect β_{T_1} and $\hat{\beta}_{OLS}$, and the second projection relates the limiting distributions of $\sqrt{T_1}(\hat{\beta}_{T_1}-\beta_0)$ and $\sqrt{T_1}(\hat{\beta}_{OLS}-\beta_0)$.

With the above definition of the projection operator Π_{Λ,T_1}, we show in Appendix A that

$$\begin{aligned}\hat{\beta}_{T_1} &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS}-\beta)'(X'X/T_1)(\hat{\beta}_{OLS}-\beta) \\ &= \arg\min_{\beta\in\Lambda}||\beta-\hat{\beta}_{OLS}||_X^2 \\ &= \Pi_{\Lambda,T_1}\hat{\beta}_{OLS}. \end{aligned} \tag{3.3}$$

Equation (3.3) says that the constrained estimator is a projection of the unconstrained estimator onto the constrained set Λ.

It is easy to check that when $X'X/T_1$ is a diagonal matrix, then there is a simple closed form solution to the constrained minimization problem (3.3). We consider a simple model without an intercept, and with two control units. For the modified synthetic control method, the constrained set is $\Lambda = \Lambda_{Msyn} = \mathcal{R}^2_+$, the first quadrant of the 2-dimensional plane. When the weight matrix is diagonal, say $X'X/T_1 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ and that $\hat{\beta}_{OLS} = (1,-1)'$. Then the objective function is $(\beta - \hat{\beta}_{OLS})'(X'X/T_1)(\beta-\hat{\beta}_{OLS}) = 2(\beta_1-\hat{\beta}_{OLS,1})^2 + 3(\beta_2-\hat{\beta}_{OLS,2})^2$, then it is easy to see that the closed form solution is $\hat{\beta}_{T_1,j} = \hat{\beta}_{OLS,j}$ if $\hat{\beta}_{OLS,j} \geq 0$; and $\hat{\beta}_{T_1,j} = 0$ if $\hat{\beta}_{OLS,j} < 0$ for $j = 1, 2$, i.e., the projection simply keeps the positive component as it is, and maps the negative component to zero. However, when $X'X/T_1$ is not a diagonal matrix, there does not exist such a simple non-iterative closed form solution. Nevertheless, we show in Appendix A that when $X'X/T_1$ is positive definite, the objective is globally convex and there is a unique solution to the constrained minimization problem. Consider a non-diagonal matrix $X'X/T_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, then for $\hat{\beta}_{OLS} = (1,-1)'$, the constrained minimizer is $(\hat{\beta}_{T_1,1}, \hat{\beta}_{T_1,2}) = (0.5, 0)$. Figure 1 shows that the objective function is a convex function and its unique minimizer occurs at $(0.5, 0)$. We know that when $\hat{\beta}_{OLS}$ lies outside the constrained set Λ, the constrained estimator $\hat{\beta}_{T_1}$ will take value at the boundary of Λ. Therefore, we can plot a 2-dimensional curve by fixing $\hat{\beta}_{T_1,2} = 0$ to see clearly that $\hat{\beta}_{T_1,1}$ takes value 0.5 as figure 2 shows. Note that when $X'X/T_1$ is non-diagonal, the projection no longer simply keeps the positive component as it is and maps negative component to zero. Or in other words, the projection does not map $(1,-1)$ to the closest point in $\Lambda = \mathcal{R}^2_+$ (which would be $(1,0)$), but due to the non-zero off-diagonal element in $X'X/T_1$, the projection maps $(1,-1)$ to $(0.5,0)$.

To derive the asymptotic distribution of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ (hence, for $\hat{\Delta}_1$), we need first to examine the asymptotic (as $T_1 \to \infty$) range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$. We will show that this asymptotic range depends on both Λ and β_0. Therefore, we will use the notation T_{Λ,β_0} to denote the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

Figure 1: The constrained LS objective function

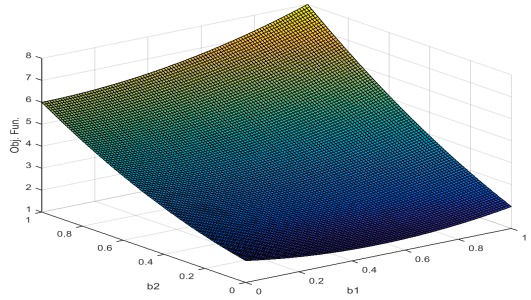
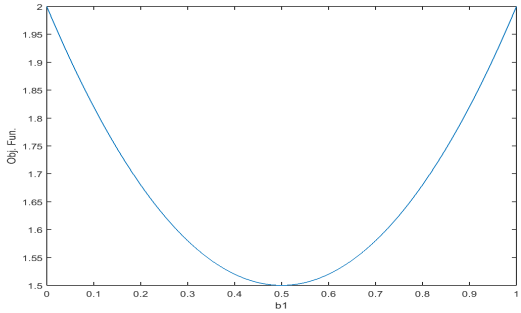


Figure 2: The objective function with $b_2 = 0$



We note that even both $\hat{\beta}_{T_1}$ and β_0 take values at the constrained set Λ, $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ does not necessarily take values in Λ, the range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ depends on Λ as well as how many components of the β_0 vector taking value 0, i.e., it depends on how many non-negativity constraints are binding. We illustrate this point via a simple example.

3.3 Examples of T_{Λ,β_0}, the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$

3.3.1 Examples of $T_{\Lambda_{Syn},\beta_0}$

We consider a simple case to illustrate the asymptotic range $T_{\Lambda_{Syn},\beta_0}$ for the standard synthetic control problem. For expositional simplicity, we consider a simple model of two control units and without an intercept: $y_{1t} = x_t'\beta_0 + u_{1t} = x_{1t}\beta_{0,1} + x_{2t}\beta_{0,2} + u_{1t}$ with $\beta_0 = (\beta_{0,1}, \beta_{0,2})' \in \mathcal{R}_+^2$ and $\beta_{0,1} + \beta_{0,2} = 1$. The constrained set can be written as $\Lambda_{Syn} = \{\beta_0 \in \mathcal{R}_+^2,\ \beta_{0,1} \in [0,1] \text{ and } \beta_{0,2} = 1 - \beta_{0,1}\}$.

To characterize $T_{\Lambda_{Syn},\beta_0}$ for $\beta_0 \in \Lambda_{Syn}$, we consider the following three inclusive cases: (i) $\beta_{0,1} = 0$ and $\beta_{0,2} = 1$; (ii) $\beta_{0,1} = 1$ and $\beta_{0,2} = 0$; (iii) $\beta_0 \in (0,1)$ and $\beta_{0,2} = 1 - \beta_{0,1}$. It is easy to check that for case (i), we have $\hat{\beta}_{T_1,1} - \beta_{0,1} = \hat{\beta}_{T_1,1} \in [0,1]$. Hence, $\sqrt{T_1}(\hat{\beta}_{T_1,1} - \beta_{0,1}) = \sqrt{T_1}\hat{\beta}_{T_1,1} \in [0, \sqrt{T_1}] \to [0,\infty) = \mathcal{R}_+$ as $T_1 \to \infty$. Similarly, we have $\sqrt{T_1}(\hat{\beta}_{T_1,2} - \beta_{0,2}) = \sqrt{T_1}(\hat{\beta}_{T_1,1} - 1) \in \sqrt{T_1}[-1,0] = [-\sqrt{T_1}, 0] \to (-\infty, 0] = \mathcal{R}_-$ as $T_1 \to \infty$. Hence, the asymptotic range of $\sqrt{T_1}(\hat{\beta} - \beta_0)$ for case (i) is $T_{\Lambda_{Syn},\beta_0,(i)} = \mathcal{R}_+ \times \mathcal{R}_-$, which is the fourth quadrant. Similarly, for case (ii), one can easily show

that $T_{\Lambda_{Syn},\beta_0,(ii)} = \mathcal{R}_- \times \mathcal{R}_+$, the second quadrant. Finally, for case (iii), since $\beta_j - \beta_{0,j}$ can be either positive or negative for $j = 1, 2$, hence, $\sqrt{T_1}(\beta_j - \beta_{0,j})$ covers the whole real line as $T_1 \to \infty$. Therefore, we have $T_{\Lambda_{Syn},\beta_0,(iii)} = \mathcal{R}^2$, the whole plane.

3.3.2 Examples of $T_{\Lambda_{Msyn},\beta_0}$

We consider the same model as discussed in section 3.3.1 except that we now consider the modified synthetic control method: $y_{1t} = x_t'\beta_0 + u_{1t} = x_{1t}\beta_{0,1} + x_{2t}\beta_{0,2} + u_{1t}$ with $\beta_0 = (\beta_{0,1}, \beta_{0,2})' \in \mathcal{R}^2_+$. To characterize $T_{\Lambda_{Msyn},\beta_0}$ for $\beta_0 \in \Lambda_{Msyn} = \mathcal{R}^2_+$, we consider four cases: (i) $\beta_0 = (0,0)'$; (ii) $\beta_0 = (0, \beta_{0,2})$ with $\beta_{0,2} > 0$; (iii) $\beta_0 = (\beta_{0,1}, 0)$ with $\beta_{0,1} > 0$; and (iv) $\beta_0 = (\beta_{0,1}, \beta_{0,2})$ with $\beta_{0,j} > 0$ for $j = 1, 2$. For case (i) we have $\sqrt{T_1}(\hat{\beta}_{T_1,j} - \beta_{0,j}) = \sqrt{T_1}\hat{\beta}_{T_1,j} \in [0, +\infty)$ for $j = 1, 2$. Hence, $T_{\Lambda_{Msyn},\beta_0,(i)} = \mathcal{R}_+ \times \mathcal{R}_+$, which is the first quadrant. For case (ii), it is easy to see that $\sqrt{T_1}(\hat{\beta}_{T_1,1} - \beta_{0,1}) = \sqrt{T_1}\hat{\beta}_{T_1,1} \in [0, +\infty)$, and $\sqrt{T_1}(\hat{\beta}_{T_1,2} - \beta_{0,2}) \in \sqrt{T_1}[-\beta_{0,2}, \infty) \to (-\infty, +\infty) = \mathcal{R}$ as $T_1 \to \infty$. Hence, $T_{\Lambda_{Msyn},\beta_0,(ii)} = \mathcal{R}_+ \times \mathcal{R}$, which is the union of the first and the fourth quadrants. Similarly, it is easy to check that $T_{\Lambda_{Msyn},\beta_0,(iii)} = \mathcal{R} \times \mathcal{R}_+$, the union of the first and the second quadrants. Finally, for case (iv), because $\hat{\beta}_{T_1,j} - \beta_{0,j}$ can be either positive or negative for $j = 1, 2$, $\sqrt{T_1}(\hat{\beta}_{T_1,j} - \beta_{0,j}) \to \mathcal{R}$ as $T_1 \to \infty$. Hence, $T_{\Lambda_{Msyn},\beta_0,(iv)} = \mathcal{R} \times \mathcal{R}$, the whole plane.

Remark 3.1 *Through the above examples one can see that T_{Λ,β_0} gives the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$. Hence, it is quite intuitive to expect that the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ can be represented as a projection of the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$ onto T_{Λ,β_0}.*

We show in the next subsection that the intuition stated in remark 3.1 is indeed correct. We present the result in the next subsection.

3.4 The asymptotic theory: the stationary data case

We term the set T_{Λ,β_0} as the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ based on intuitive argument. In the projection theory, the set T_{Λ,β_0} is termed as the tangent cone of Λ at β_0. We give a formal definition of a tangent cone as well as some explanation to the term 'tangent' in appendix B.

Theorem 3.2 *Let Z_1 denote the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$, then under the assumptions 1 to 4 presented in Appendix A, we have*

$$\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1. \tag{3.4}$$

Note that Theorem 3.2 states that the limiting distribution of the constrained estimator can be represented as a projection of the unconstrained (least squares) estimator onto the tangent cone T_{Λ,β_0}.

With the help of Theorem 3.2, we derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ as follows.

Theorem 3.3 *Under the same conditions as in Theorem 3.2, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} -\eta E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2, \tag{3.5}$$

where $\hat{\Delta}_1 = \hat{\Delta}_{1,Syn}$ *or* $\hat{\Delta}_{1,Msyn}$, $\Delta_1 = E(\Delta_{1t})$, $\eta = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, Z_1 *is defined in Theorem 3.2,* Z_2 *is independent with* Z_1 *and distributed as* $N(0,\Sigma_2)$, $\Sigma_2 = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$, $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + u_{1t}$, u_{1t} *has zero mean and is defined in (2.9).*

The proof of Theorem 3.3 is given in Appendix B.

Although one can use projection theory to characterize the asymptotic distribution of $\sqrt{T_1}(\hat{\Delta}_1 - \Delta_1)$, the inference is not straightforward as one has to know β_0 in order to calculate the tangent cone T_{Λ,β_0}. Fortunately, we will show in section 4 that a simple sub-sampling method can be used to conduct valid inference. In particular, one does not need to know β_0 when using the sub-sampling method for inferences.

3.5 The Trend-Stationary data case

Up to now we only consider the stationary data case. However, many datasets, especially for panel data with a long time dimension, exhibit some trending behaviors. For example, sales of a new product tends to go up as more people get to know the product. In this subsection we extend the stationary data result to the trend-stationary data case.

3.5.1 ATE estimation with Trend Stationary Data

From the previous analysis, we know that the synthetic control estimator is a projection of the unconstrained least squares estimator onto the constrained set Λ; and the asymptotic theory of $\sqrt{T_1}(\hat{\beta}_{Syn} - \beta_0)$ is the projection of the asymptotic distribution of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$ onto the tangent cone $T_{\lambda_{Syn},\beta_0}$. Therefore, we will first study the asymptotic distribution of the unconstrained least squares estimator using the pre-treatment data.

The trend-stationary data generating process can also be motivated using a factor model framework. Let $\{y_{it}^0\}$, for $i = 1, ..., N$ and $t = 1, ..., T$, be generated by some common factors with one of the factor being a time trend and the remaining factors being weak dependent stationary variables. Following Hsiao, Ching and Wan (2012) we assume that $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$ is generated via a factor model

$$y_t^0 = \delta_0 + Bf_t + \epsilon_t, \tag{3.6}$$

where $\delta_0 = (\delta_{01}, ..., \delta_{0N})'$ is an $N \times 1$ vector of intercepts, B is an $N \times K$ factor loading matrix, $f_t = (f_{1t}, ..., f_{Kt})'$ is a $K \times 1$ vector of common factors, $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$ is an $N \times 1$ vector of idiosyncratic errors. We assume that $f_{1t} = t$ and all other factors are stationary variables. Also, ϵ_t is a

zero mean, weakly dependent process with finite fourth moment. Hence, y_t^0 follows a trend-stationary process.

Hsiao, Ching and Wan (2012), and Li and Bell (2017) show that, under the condition that $rank(B) = K$, one can replace the unobservable factor f_t by $x_t = (1, y_{2t}, ..., y_{Nt})'$ to estimate the counterfactual outcome y_{1t}^0. Specifically, one can estimate the following regression model

$$y_{1t} = x_t'\delta + u_{1t}, \qquad (t = 1, ..., T_1), \tag{3.7}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$ and $\delta = (\delta_1, ..., \delta_N)'$.

To facility the asymptotic analysis, below we consider the time trend component explicitly. We write $y_{jt} = c_{0,j} + c_{1,j}t + y_{jt}^*$, where y_{jt}^* is a weakly dependent stationary process (de-trended from y_{jt}) for $j = 2, ..., N$. Let $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ and $\tilde{\delta} = (\delta_2, ..., \delta_N)'$. Then in vector notation, we have $\tilde{y}_t = \tilde{c}_0 + \tilde{c}_1 t + \tilde{y}_t^*$, $\tilde{c}_0 = (c_{0,2}, ..., c_{0,N})'$, $\tilde{c}_1 = (c_{1,2}, ..., c_{1,N})'$ and $\tilde{y}_t^* = (\tilde{y}_{2t}^*, ..., \tilde{y}_{Nt}^*)'$. Then we can write $\tilde{y}_t'\tilde{\delta} = (\tilde{c}_0 + \tilde{c}_1 t + \tilde{y}_t^*)'\tilde{\delta}$. Hence, we can re-write (3.7) as

$$\begin{aligned} y_{1t} &= \delta_1 + \tilde{y}_t'\tilde{\delta} + u_{1t} \\ &= \alpha_0 t + \beta_1 + \tilde{\delta}'\tilde{y}_t^* + u_{1t} \\ &= \alpha_0 t + z_t'\beta_0 + u_{1t} \qquad\qquad t = 1, ..., T_1, \end{aligned} \tag{3.8}$$

where $\alpha_0 = \tilde{c}_1'\tilde{\delta}$, $\beta_1 = \delta_1 + \tilde{c}_0'\tilde{\delta}$, $\beta_0 = (\beta_1, \tilde{\delta}')'$ and $z_t = (1, \tilde{y}^{*\prime})' \equiv (1, y_{2t}^*, ..., y_{Nt}^*)'$.

Let $\hat{\alpha}_{T_1}$ and $\hat{\beta}_{T_1}$ be the constrained least squares estimators of α_0 and β_0 subject to $\beta_j \geq 0$ for $j = 2, ..., N$ and $\sum_{j=2}^N \beta_j = 1$ for the synthetic control estimator; or dropping the sum to one restriction for the modified synthetic control estimator using the pre-treatment data. We estimate y_{1t}^0 by $\hat{y}_{1t}^0 = \hat{\alpha}_{T_1} t + z_t'\hat{\beta}_{T_1}$ and estimate the ATE is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t}, \tag{3.9}$$

where $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$.

3.5.2 Asymptotic Theory with Trend Stationary Data

To derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we need first present the theory for the unconstrained least squares estimator of $\gamma_0 = (\alpha_0, \beta_0')'$. Let $\hat{\gamma}_{OLS}$ denote the ordinary least squares estimator of γ_0. Define $M_{T_1} = \sqrt{T_1} diag(T_1, 1, ..., 1)$, which is an $(N+1) \times (N+1)$ diagonal matrix with its first diagonal element equals to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$. Then, it is well established that (e.g., Hamilton (1994))

$$M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0) \xrightarrow{d} N(0, \Omega), \tag{3.10}$$

where Ω is a $(N+1)\times(N+1)$ positive definite matrix whose explicit definition can be found in Hamilton (1994, Chapter 16).

We still use Λ to denote constrained sets for $\hat{\gamma}_{T_1}$ for trend-stationary data case. Now γ is an $(N+1)\times 1$ vector whose first component is the time trend coefficient, the second component is the intercept. Hence, the constrained sets for the standard and the modified synthetic control models are

$$\Lambda_{Syn} = \{\gamma \in \mathcal{R}^{N+1}: \ \gamma_j \geq 0 \text{ for } j=3,...,N+1, \textstyle\sum_{j=3}^{N+1}\gamma_j = 1\}; \tag{3.11}$$

$$\Lambda_{Msyn} = \{\gamma \in \mathcal{R}^{N+1}: \ \gamma_j \geq 0 \text{ for } j=3,...,N+1\}. \tag{3.12}$$

Define the synthetic control estimator

$$\hat{\gamma}_{T_1} = \arg\min_{\gamma\in\Lambda} \sum_{t=1}^{T_1}(y_{1t} - w_t'\gamma)^2, \tag{3.13}$$

where $w_t = (t, z_t')'$, $\Lambda = \Lambda_{Syn}$ or Λ_{Msyn}. Then similar to Theorem 3.2, we have

Theorem 3.4 *Let Z_3 denote the limiting distribution of $\sqrt{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in (3.10), then under the assumptions D1 to D3 presented in the supplementary Appendix D, we have*

$$\sqrt{T_1}(\hat{\gamma}_{T_1} - \gamma_0) \xrightarrow{d} \Pi_{T_{\Lambda,\gamma_0}} Z_3, \tag{3.14}$$

where T_{Λ,γ_0} is the tangent cone of Λ evaluated at γ_0, Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{OLS}-\gamma_0)$ as described in (3.10).

With Theorem 3.4 we can derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$.

Theorem 3.5 *Under the same conditions as in Theorem 3.2, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} -c'\Pi_{T_{\Lambda,\gamma_0}} Z_3 + Z_2, \tag{3.15}$$

where $\hat{\Delta}_1$ is defined in (3.9), $\Delta_1 = E(\Delta_{1t})$, $c = (\sqrt{\eta}(2+\eta), \sqrt{\eta}E(z_t'))'$, $\eta = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, Z_3 is the limiting distribution of $M_{T_1}(\hat{\gamma}_{OLS}-\gamma_0)$ as defined in (3.10), Z_2 is independent with Z_3, and it is normally distributed with zero mean and variance $\Sigma_2 = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T}E(v_{1t}v_{1s})$, $v_{it} = \Delta_{1t} - \Delta_1 + u_{1t}$.

The proof of Theorem 3.5 is similar to the proof of Theorem 3.3 and is thus omitted.

4 Inference Theory

In this section we discuss inference methods for the ATE estimator $\hat{\Delta}_1$. For expositional simplicity, we will only discuss inferences for the stationary data case. For trend-stationary data, one can first

de-trend the data, then using the inference method discussed in this section for the de-trended data. In section 4.1 we consider the case that both T_1 and T_2 are large, while in section 4.2 we deal with the small T_2 case.

4.1 Sub-sampling method

As discussed in the above section, the inference theory for the synthetic control estimator is complicated. The asymptotic distribution of $\hat{\beta}_{T_1}$ depends on whether $\beta_{0,j} = 0$ or $\beta_{0,j} > 0$ for $j = 2, ..., N$. When $\beta_{0,j} > 0$ for all $j = 2, ..., N$, asymptotically, the constraints are non-binding and the asymptotic theory of the constrained estimator is the same as that of the unconstrained ordinary least squares estimator. However, when the constraints are binding for some $j \in \{2, ..., N\}$, the asymptotic distribution of the constrained estimator is much more complex (e.g., (3.4)). The asymptotic distribution of the synthetic control coefficient estimators depends on whether the true parameters take value at the boundary or not. In practice we do not know which constraints are binding and which are not, making it more difficult to use the asymptotic theory for inference. Moreover, it is known that when parameters fall to the boundary of the parameter space, the standard bootstrap method does not work (e.g., Andrews (2000), Fang and Santos (2015)). We resolve this difficulty by proposing an easy-to-implement sub-sampling method. The proposed sub-sampling method works whether constraints are binding, partially binding or non-binding. That is, the sub-sampling method is adaptive in the sense that we do not need to know whether constraints are binding and if they are binding, we do not need to know bindings are on which coefficients.[1]

We use m to denote the sub-sample size. We will show that $\hat{\Delta}_1$ can be decomposed into two terms, one term is related to the constrained estimator $\hat{\beta}_{T_1}$, the second term is unrelated to $\hat{\beta}_{T_1}$ but depends on T_2, it can be shown that a brute-force application of the sub-sampling method will not work in general. The correct method is to first isolate the term $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ from other terms, and apply the sub-sampling method only to $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

For the whole sample period, the outcome y_{1t} is generated by

$$y_{1t} = x_t'\beta_0 + d_t\Delta_{1t} + u_{1t}, \qquad t = 1, ..., T_1, ..., T, \tag{4.1}$$

where d_t is the post-treatment time period dummy so that $d_t = 0$ if $t \leq T_1$ and $d_t = 1$ if $t \geq T_1 + 1$.

Substituting (4.1) into (3.5) we obtain

$$\hat{A} \overset{def}{=} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$$

[1]Hong and Li (2017) show that numerical differentiation bootstrap method can consistently estimate the limiting distribution in many cases where the conventional bootstrap is known to fail. One can also use Hong and Li's (2017) method to conduct inference for the synthetic control estimator. In this paper we focus on the simple sub-sampling method.

$$
\begin{aligned}
&= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0 - \Delta_1) \\
&= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (x_t'\beta_0 + \Delta_{1t} + u_{1t} - x_t'\hat{\beta}_{T_1} - \Delta_1) \\
&= -\sqrt{\frac{T_2}{T_1}} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T} x_t' \right] \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (\Delta_{1t} - \Delta_1 + u_{1t}) \\
&\equiv -\sqrt{\frac{T_2}{T_1}} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T} x_t' \right] \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t}, \qquad (4.2)
\end{aligned}
$$

where $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$.

Now we impose an additional assumption that u_{1t} and v_{1t} are both serially uncorrelated. This assumption greatly simplifies the sub-sampling method that will be discussed below. This assumption can be tested easily in practice. When this assumption is violated, more sophisticated method such as block sub-sampling method can be used to deliver valid inferences.

Expression (4.2) suggests that we only need to apply the sub-sampling method to the term $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ because only this term is related to the constrained estimator. We now describe the sub-sampling steps. In Appendix B we show that, when v_{1t} is serially uncorrelated, one can consistently estimate Σ_2 by $\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} \hat{v}_{1t}^2$, where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. We generate $v_{1t}^* \sim$ iid $N(0, \hat{\Sigma}_2)$ for $t = T_1 + 1, ..., T$. Next, let m be the sub-sample size that satisfies the condition that $m \to \infty$ and $m/T_1 \to 0$ as $T_1 \to \infty$. For $t = 1, ..., m$, we randomly draw (y_{1t}^*, x_t^*) from $\{y_{1t}, x_t\}_{t=1}^{T_1}$ without replacement (sub-sampling). Then we use the sub-sample $\{y_{1t}^*, x_t^*\}_{t=1}^{m}$ to estimate β_0 by the constrained least squares method, i.e., $\hat{\beta}_m^* = \arg\min_{\beta\in\Lambda} \sum_{t=1}^{m}(y_{1t}^* - x_t^{*'}\beta)^2$. The sub-sampling/bootstrap version of the statistic $\hat{A}$ is given by

$$
\hat{A}^* = -\sqrt{\frac{T_2}{T_1}} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T} x_t' \right] \sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t}^*. \qquad (4.3)
$$

We repeat the above process for a large number of times, say, J times. Using $\{\hat{A}_j^*\}_{j=1}^{J}$, we can obtain confidence intervals for $\hat{A}$.

Let $\hat{A}_{(1)}^* \le \hat{A}_{(2)}^* \le ... \le \hat{A}_{(J)}^*$. Then the $1-\alpha$ confidence interval for Δ_1 is given by

$$
[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{(1-\alpha/2)}^*,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{(\alpha/2)}^*]. \qquad (4.4)
$$

We show that the above method indeed gives consistent estimation of the confidence intervals for Δ_1 in the next Theorem.

Theorem 4.1 *Under the same conditions as in Theorem 3.3, also assuming that $m \to \infty$ and $m/T_1 \to 0$ as $T_1 \to \infty$, then for any $\alpha \in (0,1)$, the $(1-\alpha)$ confidence interval of Δ_1 can be consistently estimated*

by $[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(1-\alpha/2)}\,,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(\alpha/2)}]$.

The sub-sampling method is a powerful tool for inference. It works under quite general conditions even when the regular bootstrap method does not work as in the case of the synthetic control ATE estimator. Politis, Romano and Wolf (1999)) provide proofs/arguments showing that 'sub-sampling method works' under very weak regularity conditions.

Remark 4.2 *Note that even we random draw* (y_t^*, x_t^*) *from* $\{y_s, x_s\}_{s=1}^{T_1}$ *for* $t = 1, ..., m$*, we do not require that* $\{y_s, x_s\}_{s=1}^{T_1}$ *to be a serially uncorrelated process. In fact, they can have arbitrary serial correlation, say* $\{y_{jt}\}_{j=1}^{N}$ *is generated by some serially correlated common factors, we only need that idiosyncratic error* u_{1t} *in (2.9) is serially uncorrelated. This can be easily tested given data. In section 5 we generate* y_{jt} *using a three factor model, where the three factors follow AR, ARMA and MA processes, respectively. Simulations show that the above proposed sub-sampling method works well. When* u_{1t} *is serially correlated, we conjecture that one replace the random sub-sampling method by block sub-sampling method. We leave the formal justification of using block sub-sampling method as a future research topic.*

Remark 4.3 *In the literature of sub-sampling method, the choice of sub-sample size* m *is a key issue. Bickel and Sakov (2008) propose a data-driven method to select* m*. In general, too small an* m *or too large an* m *do not work well, when* m *falls into an appropriate interval, the performance should be stable and acceptable. For our model, because* $\beta_{0,j} > 0$ *for some* $j \in \{2, ..., N\}$*, and that the statistic*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}, \tag{4.5}$$

also contains a term $T_2^{-1}\sum_{t=T_1+1}^{T} v_{1t}$*, where* $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$*, which is not related to* $\hat{\beta}_{T_1}$*. It turns out that the sub-sampling method works well for a wider range of* m*. We will discuss more on this issue at Section 5.3 and at the supplementary Appendix D.*

4.2 Inference theory when T_2 is small

The asymptotic theories presented in section 4.1 assume that both T_1 and T_2 are large. However, in practice, many datasets have T_2 much smaller than T_1. When T_2 is small, Ferman and Pinto (2016a) propose using Andrews' (2003) end-of-sample instability test to conduct inference for DID and synthetic control ATE estimators. In this subsection we discuss using Andrews's test to conduct inferences for the ATE estimator based on the synthetic control method.

When T_1 is large and T_2 is small, the first term on the right-hand-side of (4.2) has an order

$O_p(\sqrt{T_2/T_1}) = O_p(T_1^{-1/2})$ becomes negligible, then we have

$$\hat{A} = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0 - \Delta_1) = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1), \tag{4.6}$$

where $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$ has zero mean and finite variance.

One can test the null hypothesis of a constant treatment effects H_0: $\Delta_{1t} = \Delta_{1,0}$ for some pre-specified value $\Delta_{1,0}$ for $t = T_1 + 1, ..., T$, against, say, a one-sided positive treatment effects H_1: $\Delta_1 = E(\Delta_{1t}) > \Delta_{1,0}$ for $t = T_1 + 1, ..., T$. Following Andrews (2003), we can use the following test statistic

$$\hat{B}_{T_2} = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0 - \Delta_{1,0}) = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t,0} + o_p(1), \tag{4.7}$$

where $v_{1t,0} = \Delta_{1t} - \Delta_{1,0} + u_{1t}$. Under H_0 we have $v_{1t,0} = u_{1t}$ has zero mean, and it has a positive mean under H_1.

To conduct inference based on the test statistic $\hat{B}_{T_2}$, we compute the following quantity

$$\hat{B}_{T_2,j} \overset{def}{=} \frac{1}{\sqrt{T_2}} \sum_{t=j}^{T_2+j-1} \hat{u}_{1t} = \frac{1}{\sqrt{T_2}} \sum_{t=j}^{T_2+j-1} u_{1t} + O_p(T_1^{-1/2}), \qquad \text{for } j = 1, ..., T_1 + 1 - T_2, \tag{4.8}$$

where for $t = 1, ..., T_1$, $\hat{u}_{1t} = y_{1t} - \hat{y}_{1t}^0 = x_t'(\beta_0 - \hat{\beta}_{T_1}) + u_{1t} = u_{1t} + O_p(T_1^{-1/2})$ because $\hat{\beta}_{T_1} - \beta_0 = O_p(T_1^{-1/2})$. The empirical distribution of $\{\hat{B}_{2,j}\}_{j=1}^{T_1+1-T_2}$ can be used to obtain critical values for the test statistic $\hat{B}_{T_2}$ under the null hypothesis H_0: $\Delta_{1t} = \Delta_{1,0}$ for all $t = T_1 + 1, ..., T$. If $\hat{B}_{T_2}$ is at the tail of this empirical distribution, we reject the null hypothesis and accept the alternative hypothesis.

Remark 4.4 *We can only test a constant treatment effects for each post-treatment period using Andrews' test, i.e., we can only test $\Delta_{1,t} = \Delta_{1,0}$ for all $t = T_1 + 1, ..., T$. We cannot test $\Delta_1 = \Delta_{1,0}$ because under this null hypothesis, we will have $\Delta_{1,t} - \Delta_1 = \Delta_{1t} - \Delta_{1,0}$ has a zero mean and a finite variance, we cannot use pre-treatment data to estimate the variance of Δ_{1t}. Therefore, Andrews' method become invalid when treatment effects varies with t.*

Remark 4.5 *Andrews' test will have good estimated sizes for large T_1. However, it is not a consistent test because T_2 is small. The power of the test depends on the strength of the treatment effects under H_1. The power of the test should increase with T_2, but when T_2 is large, an estimation error of order $\sqrt{T_2/T_1}$ may become non-negligible, rendering Andrews' test invalid in our context, see section 5.3 for a more detailed discussion on this issue.*

5 Monte Carlo simulations

In this section we first consider the case of large T_1 and T_2 and examine the performance of sub-sampling method inferences through simulations, then we consider the case of large T_1 and small T_2

and examine the performance of Andrews' (2003) end-of-sample instability test.

5.1 A three factor data generating process

We conduct simulation studies using the same data generating process as in Hsiao, Ching and Wan (2012) and Du and Zhang (2015). We consider the same 3-factor model as follows.

$$\begin{aligned} f_{1t} &= 0.8f_{1t-1} + \epsilon_{1t}, \\ f_{2t} &= -0.6f_{1t-1} + \epsilon_{2t} + 0.8\epsilon_{2t-1}, \\ f_{3t} &= \epsilon_{3t} + 0.9\epsilon_{3t-1} + 0.4\epsilon_{3t-2}, \end{aligned}$$

where ϵ_{it} is iid $N(0,1)$. Let y_t^0 denote the $N \times 1$ vector of outcome variables without treatment. It is generated via the factor model

$$y_t^0 = a + Bf_t + u_t, \qquad t = 1, ..., T \tag{5.1}$$

where $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$, $a = (a_1, a_2, ..., a_N)'$ and $u_t = (u_{1t}, u_{2t}, ..., u_{Nt})'$ are all $N \times 1$ vectors, $B = (b_1, b_2, ..., b_N)'$ is the $N \times 3$ loading matrix where b_j is a 3×1 loading vector for unit j, f_t is the 3×1 common factors: $f_t = (f_{1t}, f_{2t}, f_{3t})'$. We choose $(a_1, a_2, ..., a_N) = (1, 1, ..., 1)$, ϵ_{jt} iid $N(0, \sigma^2)$ with $\sigma^2 = 0.5$.

We use a set up similar to our Warby Parker empirical data by setting $T_1 = 90$, $T_2 = 20$, $T = T_1 + T_2 = 110$ and $N = 11$ (with 10 control units). For factor loadings, we use b_1 to denote the 3×1 vector of loadings for the first unit (the treated unit); use $\tilde{b}_2 = (b_2, ..., b_{s+1})$ denote $3 \times s$ loading matrix for units $j = 2, ..., s+1$ $(1 \leq s \leq N-2)$; finally we use $\tilde{b}_3 = (b_{s+1}, ..., b_N)$ denote the $3 \times (N-1-s)$ loading matrix for the last $N-s-1$ units. We fix $s = 6$ and consider the following two sets for factor loadings:

$$\begin{aligned} DGP1: &\quad b_1 = \mathbf{1}_{3\times1}; \quad b_j = \mathbf{1}_{3\times1} \text{ for } j = 2, ..., 7; \quad \text{and} \quad b_j = \mathbf{0}_{3\times1} \text{ for } j = 8, ..., 11, \\ DGP2: &\quad b_1 = 2(\mathbf{1}_{3\times1}); \quad b_j = \mathbf{1}_{3\times1} \text{ for } j = 2, ..., 7; \quad \text{and} \quad b_j = \mathbf{0}_{3\times1} \text{ for } j = 8, ..., 11, \end{aligned}$$

where $\mathbf{1}_{3\times1}$ and $\mathbf{0}_{3\times1}$ denote 3×1 vectors of ones and zeros, respectively.

Note that for both DGP1 and DGP2, 6 out of 10 control units have non-zero loadings and the remaining 4 control units have zero loadings. Also note that for DGP1, all non-zero factor loadings are set to be ones so that the treated and the control units (with non-zero loadings) are drawn from a common distribution. While for DGP2, loadings for the treated unit all equal to 2, and the controls units' loadings (with non-zero loadings) are all equal to 1. Thus, the treated and control units are drawn from two heterogeneous distributions.

We generate the following treatment effects Δ_{1t}:

$$\Delta_{1t} = \alpha_0 \left[\frac{e^{z_t}}{1+e^{z_t}} + 1 \right], \qquad t = T_1 + 1, ..., T, \tag{5.2}$$

where $z_t = 0.5 z_{t-1} + \eta_t$ and η_t is iid $N(0, 0.5^2)$.

Note that for post-treatment period, $y_{1t} = y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$, where y_{1t}^0 are generated as described earlier and Δ_{1t} is generated by (5.2). There is a zero, or a positive treatment effects corresponding to $\alpha_0 = 0$ and $\alpha_0 > 0$, respectively.

5.2 Simulations results for coverage probabilities

In this section we report estimated coverage probabilities. Since we have $N = 11$ parameters in the regression model, we need to select sub-sample size $m > N$. We select $m = 20$, 40, 60, 80 and 90. Note that $T_1 = 90$ so that we include the case where the sub-sample size m equals the full sample size. The reason for considering $m = T_1$ was discussed in remark 4.3.

Table 1 reports estimated coverage probabilities for DGP1 based on 1000 simulations, and 400 sub-samplings within each simulation. The top panel corresponds to no treatment effects ($\alpha_0 = 0$) while the bottom panel corresponds to the case of a positive treatment effects with $\alpha_0 = 1$ in (5.2). From Table 1 we observe that both the standard synthetic control and the modified synthetic control methods give estimated coverage probabilities that are close to their nominal levels. Also, we observe that the estimation results are not sensitive to different values of α_0 (the magnitude of the treatment effects).

Table 1: Coverage probabilities for DGP1 (a common distribution)

	DGP1 with $\alpha_0 = 0$ (zero treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.499	0.492	0.462	0.500	0.482	0.517	0.489	0.488	0.507	0.493
80%	0.767	0.786	0.762	0.788	0.778	0.785	0.798	0.786	0.800	0.790
90%	0.883	0.890	0.879	0.889	0.885	0.894	0.879	0.882	0.885	0.883
95%	0.940	0.934	0.940	0.945	0.936	0.942	0.945	0.940	0.945	0.938
	DGP1 with $\alpha_0 = 1$ (positive treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.497	0.510	0.509	0.466	0.483	0.497	0.510	0.509	0.466	0.483
80%	0.805	0.775	0.784	0.778	0.782	0.805	0.775	0.784	0.778	0.782
90%	0.903	0.868	0.891	0.877	0.884	0.903	0.868	0.891	0.877	0.884
95%	0.944	0.931	0.950	0.929	0.934	0.944	0.931	0.950	0.929	0.934

Table 2 reports estimated coverage probabilities for DGP2 based on 1000 simulation, and 400 sub-samplings within each simulation. From Table 2 we observe that the standard synthetic control

method give biased estimation results. The estimated coverage probabilities are much smaller than the corresponding nominal values. The reason for this biased estimation result is that for DGP2, the treated and the control units are drawn from different distributions because they have different factor loadings. In contrast to the standard synthetic control approach, the modified synthetic control method give good estimated coverage probabilities. This verifies that the modified synthetic control method is robust to data drawn from heterogeneous distributions. Like the case of DGP1, the results are not sensitive to different values of α_0.

Table 2: Coverage probabilities for DGP2 (a heterogenous distribution)

	DGP2 with $\alpha_0 = 0$ (zero treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.294	0.308	0.314	0.292	0.306	0.474	0.458	0.492	0.474	0.470
80%	0.526	0.534	0.522	0.510	0.540	0.776	0.756	0.770	0.742	0.738
90%	0.658	0.630	0.638	0.632	0.666	0.884	0.854	0.876	0.844	0.866
95%	0.752	0.710	0.720	0.720	0.754	0.936	0.924	0.930	0.908	0.926
	DGP2 with $\alpha_0 = 1$ (positive treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.306	0.278	0.276	0.278	0.286	0.508	0.486	0.468	0.478	0.482
80%	0.522	0.478	0.510	0.472	0.496	0.802	0.764	0.796	0.796	0.770
90%	0.634	0.614	0.620	0.580	0.594	0.888	0.890	0.894	0.894	0.884
95%	0.710	0.716	0.710	0.678	0.668	0.948	0.944	0.940	0.950	0.944

5.3 Inferences when T_2 is small

In this section we consider case of large $T_1 = 100$, 200 and small $T_2 = 3$, 5. We use Andrews' (2003) end-of-sample instability test discussed in section 4.2 to test the null hypothesis H_0: $\Delta_{1t} = 0$ ($\Delta_{1,0} = 0$) against the one-sided alterative H_1: $\Delta_{1t} > 0$ for all $t = T_1+1, ..., T$. The data is generated by the three factor model (DGP1) as discussed in section 5.1, and the treatment effects is generated via (5.2) with $\alpha_0 = 0$ under H_0, and $\alpha_0 = 0.5$, 1 under H_1. The number of simulation is 10,000. The simulations results are reported in Table 3.

Andrews' (2003) test is expected to give good estimated sizes when T_1 is large. As expected, we see from Table 3 that the test is over sized for $T_1 = 100$, its estimated sizes improve as T_1 increase to 200. Another result worth noticing from Table 3 is that, if we fix T_1, the estimated sizes deteriorate as T_2 increases. That is understandable because this test is designed for large T_1 and small T_2.

As Andrews (2003) pointed out, this statistic is not a consistent test for small values of T_2.[2] Note that a large T_1 helps to give better estimated sizes, it does not increase the power of the test. Therefore,

[2] A test is said to be a consistent test if, when the null hypothesis is false, the probability of rejecting the (false) null hypothesis converges to one as sample size goes to infinity ($T_2 \to \infty$).

Table 3: Coverage probabilities for DGP1 (Andrews' instability test)

	H_0: $\alpha_0 = 0$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.0849	0.1362	0.2366	0.0935	0.1497	0.2440
200	0.0652	0.1161	0.2191	0.0711	0.1250	0.2273
	H_1: $\alpha_0 = 0.5$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.2892	0.4076	0.6656	0.3492	0.4753	0.6985
	H_1: $\alpha_0 = 1$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.5416	0.6573	0.7937	0.6994	0.7939	0.8853

we only consider $T_1 = 100$ for power calculation because for $T_1 = 200$ or even larger T_1, the powers of the test are similar. When T_1 is large, the power of the test increases with T_2 and also depends on the magnitude of $\sum_{t=T_1+1}^{T}(\Delta_{1t} - \Delta_{1,0})$ under H_1. From Table 3 we see that the estimated power increases with T_2 as well as with α_0 (the magnitude of Δ_{1t}). However, a large T_2 adversely affects the estimated sizes of Andrews' test.

We also conducted simulations of Andrews' test under DGP1 using $T_1 = 90$ and $T_2 = 20$ (same T_1 and T_2 as in our empirical data). Based on 10,000 simulations with $\alpha_0 = 0$, the estimated sizes are 0.1660, 0.1964 and 0.2699 for nominal levels 5%, 10%, and 20%, respectively. We see that for $T_2 = 20$, $T_1 = 90$ is not large enough for the test to have good estimated sizes, because an error term of order $\sqrt{T_2/T_1}$ is not negligible which causes Andrews' test invalid in our context. Therefore, the end-of-sample stability testing and the sub-sampling testing procedures are complement to each other. The former can be used when T_2 is small, while the later is preferred when T_2 is not small.

Remark 5.1 *Here for our synthetic control ATE estimator with panel data, large T_2 invalidates Andrews' test due an error term of order $\sqrt{T_2/T_1}$ becoming non-negligible. This differs from the time series model considered by Andrews (2003), where when T_2 is also large, testing a possible structure break at T_1 becomes a simple and standard problem.*

6 An Empirical Application

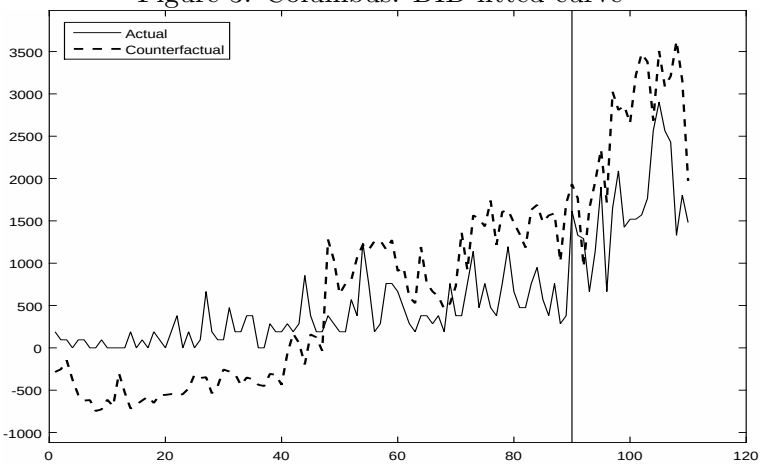
In this section we present an empirical application to illustrate the usefulness of the modified synthetic control method in practice. In the application, we calculate the ATE based on the modified synthetic control method and the confidence intervals using the sub-sampling method.

6.1 Warby Parker and ATE estimation

We have data from WarbyParker.com whose mission is to provide high quality products at lower prices ($95 instead of $300+range). In February 2010, WarbyParker.com opened its first physical showroom at the New York City headquarter. Later, they opened more showrooms in several cities hoping that opening physical showrooms can significantly promote sales. They opened a showroom in Columbus, Ohio on October 11, 2011. In this section, we want to evaluate how the showroom opening in Columbus affects Columbus' average weekly sales (the average treatment effect) in the post-treatment period. As discussed in section 2 on estimating treatment effects using panel data, we estimate the counterfactual sales for Columbus by letting the sales of Columbus be the dependent variable, and the control cities' sales (sales in cities without showrooms) be the explanatory variables. Hence, we run a constrained regression, i.e., we regress weekly sales of Columbus on sales of control cities to obtain the estimated coefficients under the restriction that the coefficients are non-negative. Then, using these estimated coefficients, together with the post-treatment period sales for the control group cities, we compute the counterfactual of what sales would be for Columbus in the absence of the showroom opening. The 10 largest cities (by population) that do not have showrooms were selected as the control group cities. These cities are: Atlanta, Chicago, Dallas, Denver, Houston, Minneapolis, Portland, San Diego, Seattle and Washington.

First, we would like to show that "parallel lines" assumption is violated for this data. Hence, DID and the standard synthetic control methods should not be used to estimate ATE for this data.

Figure 3: Columbus: DID fitted curve



In figure 3, the solid line denotes Columbus' weekly sales (in dollars) while the dashed line is fitted curve using the DID method which is computed using (2.8). The vertical line denotes the time of showroom opening (occurred at $T_1 = 90^{th}$ week). We see that the DID fitted curve and the real data have different upward trends. This is a case that DID method should not be used to estimate ATE. However, if one mis-applies the DID method to estimate ATE, one would obtain a severely biased

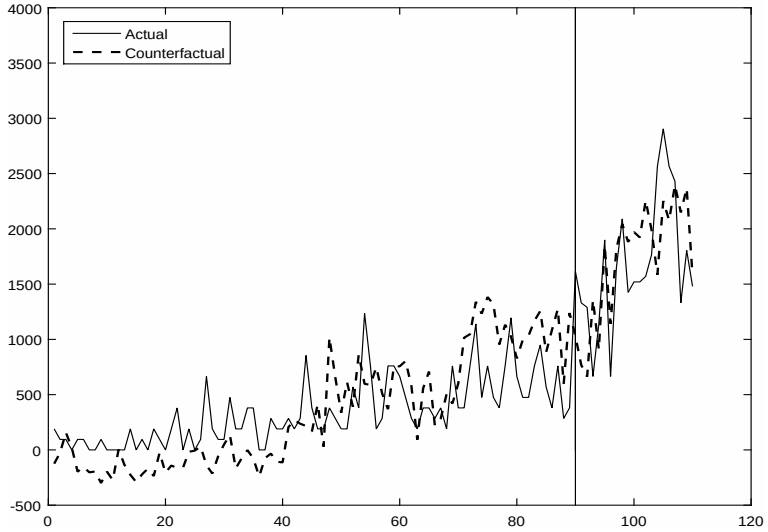
estimation result. The DID fitted curve underestimates Columbus' sales for the first half sample (from $t = 1$ to about $t = T_1/2$) and overestimates the sales for the second half sample (from $t = T_1/2 + 1$ to $t = T_1$). As a result, DID would overestimate the out-of-sample counterfactual outcome (for $t \geq T_1 + 1$ with $T_2 = 20$). Consequently, DID would underestimate the average treatment effects.

The problem with the DID method when applied to Columbus' data is that the average of the 10 control cities' sale sand Columbus' sales exhibit different upward trends during the pre-treatment period data (in the absence of treatment). Therefore, the simple average of these control cities' sales predicted Columbus' counterfactual sales poorly. The validity of the DID method relies on the assumption that the Columbus' sales and the average of control cities' sales follow parallel paths in the absence of treatment (for pre-treatment period). This assumption is violated for the data we have.

Next, we examine the standard synthetic control method. Figure 4 plots Columbus' actual sales (solid line) and the in-sample-fit and out-of-sample counterfactual forecast (dotted line) curve computed using (2.12). From figure 4, we see that the synthetic control method's in-sample-fit is also poor as it underestimates the actual sales for the first half of the in-sample data and overestimates the actual sales for the second half of the in-sample data. In this case, one should not use the standard synthetic control method to estimate ATE. Nevertheless, we can also see that if one mis-applies the synthetic control method to this data set, one would overestimate the counterfactual outcome which would result in an underestimation of ATE. The reason for this is that without the restriction that coefficients-add-to-one, the sum of the slope coefficients is 0.234. The standard synthetic control method imposes the slope coefficients to add to one, which inflates the slope of fitted curve to be larger than the slope of the actual data. The estimated intercept moves the fitted curve down parallel in an attempt to make the fitted curve and the actual data have the same sample mean (for pre-treatment period data). This leads to the fitted curve that is below the actual data for the first half of the sample data and above the actual data for the second half of the sample data. Hence, it leads to a significant overestimation of the out-of-sample counterfactual sales, which in turn leads to a severely downward biased estimated ATE.

The above analysis suggests that restricting the slope coefficients to add to one is the reason for a large estimation bias of the standard synthetic control method. Therefore, we relax the weights add-to-one condition, i.e., we only keep the non-negativity of the weights but drop the add-to-one restriction. Applying this 'modified synthetic control method' to Columbus' data, we obtain that the estimated weights add up to 0.234 which is substantially less than one. The estimation results are plotted in Figure 5. The results in figure 5 show a much improved in-sample-fit. Unlike Figures 3 and 4, the fitted curve in figure 5 does not appear to have any systematic estimation bias (for $1 \leq t \leq T_1$). Our estimation result shows that opening a showroom in Columbus on November 10, 2011 leads to an average 67% increase in weekly sales. In the next subsection we show that the estimated positive ATE is statistically highly significant.

Figure 4: Columbus: The synthetic control fitted curve



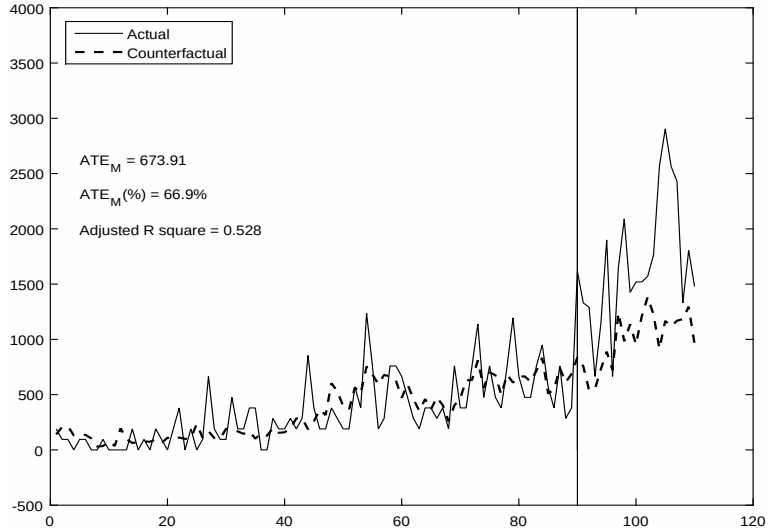
6.2 Confidence intervals for the ATE

In this section we use the sub-sampling method discussed in Section 4 to estimate confidence intervals (CI) for the ATE (Δ_1). Since our proposed sub-sampling method requires that the idiosyncratic error u_{1t} defined in (2.9) be serially uncorrelated, we will first test whether this assumption holds true for the data we have. Let $\hat{u}_{1t} = y_{1t} - x_t'\hat{\beta}_{T_1}$, then a simple statistic for testing zero serial correlation is $\sqrt{T_1}\hat{\rho} = \sqrt{T_1}\sum_{t=2}^{T_1}\hat{u}_{1t}\hat{u}_{1,t-1}/\sum_{t=2}^{T_1}\hat{u}_{1t}^2$. Under the null hypothesis of zero serial correlation, $\sqrt{T_1}\hat{\rho}$ has a standard normal distribution (for large T_1). For Columbus sales' data, we obtain $\sqrt{T_1}\hat{\rho} = 0.7281$, which gives a p-value of 0.467. Therefore, the empirical data supports the null hypothesis that there is no zero serial correlation in the idiosyncratic error u_{1t}, justifying the use of the simple sub-sampling method discussed in section 4.

To conduct the sub-sampling inference, we choose sub-sample sizes $m = 20, 40, 60, 80, 90$. For each value of m, we conduct 10,000 sub-sampling simulations $\{\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_j^*\}_{j=1}^{10,000}$. We then sort these 10,000 statistics to obtain $\alpha/2$ percentile and $(1-\alpha/2)$ percentile for $\alpha = 0.2$, 0.1, 0.05 and 0.01. The results are given in Table 4.

First, we observe that the estimated confidence intervals are quite similar for different sub-sample sizes, including the case of $m = T_1$ (recall that $T_1 = 90$). The empirical data example further verifies that, due to the reason discussed in remark 4.3, the sub-sampling method works well for a wide range of m values. Next, we notice that the lower bound of these intervals are all positive and far above zero for all m values. This implies that the estimated ATE value of 674.5 is positive and significantly different from zero for all conventional significant levels. In fact, if we conduct a 1% level test, we will reject $\Delta_1 = 295$ and in favor of $\Delta_1 > 295$ because 99% CIs are at the right of 295 for all m values considered. Similarly, if we conduct a 10% level test, we will reject $\Delta_1 = 430$ and in favor of $\Delta_1 > 430$ because all 90% CIs are at the right of 430. Thus, opening a showroom at Columbus significantly

Figure 5: Columbus: Modified synthetic control ATE estimation



increases WarbyParker.com's eyeglasses sales.

Table 4: Confidence intervals (based on 10,000 simulations)

	m=20	m=40	m=60	m=80	m=90
80% CI	[489.6, 880,1]	[487.4, 870.1]	[491.7, 876.5]	[487.8, 871.4]	[488.4, 876.5]
90% CI	[436.3, 941.9]	[431.5, 927.8]	[432.9, 926.4]	[437.5, 921.6]	[433.9, 929.9]
95% CI	[395.1, 996.0]	[389.6, 975.5]	[390.9, 978.4]	[392.2, 967.6]	[387.4, 977.6]
99% CI	[295.6, 1110.1]	[309.8, 1068.1]	[299.0, 1074.1]	[302.1, 1069.0]	[297.6,1079.5]

6.3 Robustness Checks

In this section we conduct the following robustness checks:

1. Change the treatment date from $T_1 = 90$ to a pseudo treatment date $T_0 = T_1 - 10 = 80$.
2. Add three covariates (monthly data linear interpolated to weekly data): Unemployment rate, Labor force and Average weekly earnings (weekly average income in dollars) for all employees in private sector.
3. Selecting control units based on covariates matching.
4. Comparison with the unconstrained (least squares) estimation method.

6.3.1 Change the treatment date

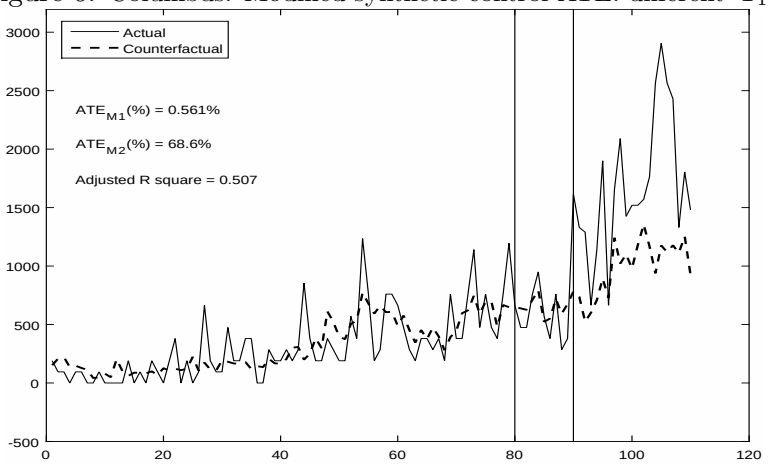
Columbus opened showroom at $t = 90$ ($T_1 = 90$). We change the treatment date to be 10 weeks earlier so it is as if the showroom were opened at $t = 80$. Using data from $t = 1$ to 80 we estimate

the model using the modified synthetic control method, then we predict Columbus' counterfactual sale from weeks 81 to 110. Since there were no showroom during $t = 81$ to 90, there should not be significant differences between y_{1t} and $\hat{y}_{1t}^0$ for $81 \leq t \leq 90$. From figure 6 we see that for the period 81 to 90, the predicted sale traces the actually sale quite closely. The percentage increase of ATE for this 10 period is 0.561% which is quite close to no effect as expected, while ATE for $t = 91$ to 110 is 68.6% which is very close to the original ATE estimate is 67%. We also compute the 80%, 90%, 95% and 99% confidence intervals (CIs) for Δ_1 based on estimation result $\hat{\Delta}_1$ using data from $t = 81$ to 90 with 10,000 sub-sampling simulations. The results are given in Table 5. We see that all CIs contain zero. Hence, we cannot reject the null hypothesis that there is no treatment effects during the period of $81 \leq t \leq 90$ at any conventional levels. Thus, this robust check supports the modified synthetic control estimation result.

Table 5: Confidence intervals (based on 10,000 simulations)

	m=20	m=40	m=60	m=80	m=90
80% CI	[-132.7,196.1]	[-136.2,176.6]	[-136.3,169.9]	[-137.9,165.5]	[-138.1,162.7]
90% CI	[-176.4,251.5]	[-176.0,224.4]	[-178.8,216.5]	[-178.1,210.3]	[-181.3,204.5]
95% CI	[-214.8,308.1]	[-213.9,267.1]	[-216.2,255.0]	[-215.5,251.5]	[-215.7,242,4]
99% CI	[-284.6,454.2]	[-295.6,354.1]	[-276.7,340.9]	[-290.3,333.7]	[-289.7,318.3]

Figure 6: Columbus: Modified synthetic control ATE: different 'T_1'



6.3.2 Adding Covariates

We collect monthly data on unemployment rate (Unemp), labor force (LF) and average weekly earnings (Inc) for Columbus, and linear extrapolate them to weekly data. The data is downloaded from the

Bureau of Labor Statistics website (bls.gov). The estimation model is

$$y_{1t} = x_t'\beta_0 + z_{1t}'\gamma_0 + u_{1t}, \qquad t = 1, ..., T_1 \tag{6.1}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$, $z_{1t} = (Unemp_t, LF_t, Inc_t)'$, β_0 and γ_0 are $N \times 1$ and 3×1 vector of parameters, respectively. Since obviously that opening a showroom has no (or negligible) effects on z_{1t}, we can use the above model to predict post-treatment counterfactual sale for the treated city. Specifically, we estimate model (6.1) under the restriction $\beta_j \geq 0$ for $j \geq 2$ using the pre-treatment data $t = 1, ..., T_1$ (there is no restriction for other parameters). Let $\hat{\beta}_{T_1}$ and $\hat{\gamma}_{T_1}$ denote the corresponding estimators. We estimate the counterfactual outcome y_{1t}^0 by $\hat{y}_{1t}^0 = x_t'\hat{\beta}_{T_1} + z_{1t}'\hat{\gamma}_{T_1}$ for $t = T_1 + 1, ..., T$ and estimate ATE by $T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t}^0)$.

Figure 7: Columbus: Modified synthetic control ATE, add Covariates

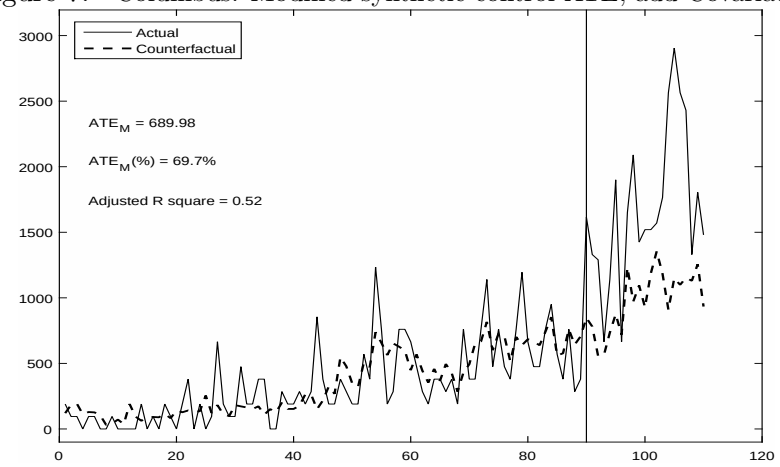


Figure 7 plots the estimation result for Columbus. The ATE becomes 68.7% which is quite close to the original result of 67%. However, the adjusted R^2 decreased slightly from 0.528 to 0.524, indicating that the three covariates do not have additional explanatory power to explain sale. The virtually same ATE estimation result with added covariates again supports our original ATE estimation result.

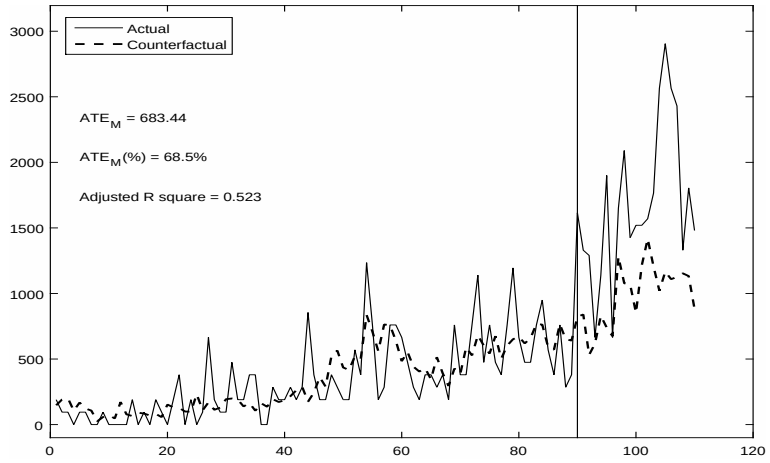
6.3.3 Select control units based on covariate matching

In this subsection we first select cities whose covariates are close to the covariates of the treated city. Then we select the number of control cities by comparing adjusted R^2. Finally we estimate ATE using the selected control units. We explain this procedure in more details below.

For each $j = 1, 2, 3$ (corresponding to Unemp, LF, Inc), we regress $z_{1,jt}$ on $z_{i,jt}$ using the pre-treatment data and obtain the goodness-of-fit $R_{i,j}^2$ for $i = 2, ..., 11$. We obtain a total R-square for city i by $R_i^2 = R_{i,1}^2 + R_{i,2}^2 + R_{i,3}^2$. We order them in a non-increasing order: $R_{(2)}^2 \geq R_{(3)}^2 \geq ... \geq R_{(11)}^2$. Their corresponding sales are denoted by $y_{(2),t}$,...,$y_{(11),t}$ for $t = 1, ..., T_1$. Next, we regress y_{1t} on $y_{(2),t}$ and obtain an adjusted $\bar{R}_{(2)}^2$; and we regress y_{1t} on $(y_{(2),t}, y_{(3),t})$ and obtain an adjusted $\bar{R}_{(2),(3)}^2$; continuing

this way until we regress y_{1t} on all $(y_{(2),t}, ..., y_{(11),t})$. We choose a model with the largest adjusted $\bar{R}^2$. For Columbus, this method selects seven cities (Portland, Houston and Atlanta are not selected) gives the largest adjusted $\bar{R}^2$. Using the seven selected cities as control group, the modified synthetic control method's estimation result is plotted in figure 8. The ATE estimation result is 68.5% which is quite close to the original result of 67%.

Figure 8: Columbus: ATE Estimation Based on Covariates Matching



6.3.4 Comparison with with the unconstrained estimator

In this subsection we consider using the ordinary least squares method to estimate the counterfactual outcome. Let $\hat{\beta}_{OLS}$ denote the least squares estimator of β using the pre-treatment sample, then the counterfactual outcome is estimated by $\hat{y}_t^0 = x_t'\hat{\beta}_{OLS}$ (e.g., Hsiao, Ching and Wan (2012)). Applying this method to the Columbus data gives an estimated ATE of \$645.26 increase in weekly sale after the opening of a showroom in Columbus. While this number is close to the ATE estimation result of \$673.72 by the modified synthetic control, we would like to compare the out-of-sample forecasting performances of the two estimation method in order to judge which method gives a more accurate ATE estimation result.

The difference between the least squares method and our modified synthetic control method is that the least squares method does impose non-negativity restriction on the slope coefficients when estimating the regression model with the pre-treatment data. The rationale for imposing the non-negativity constraints is that outcome variables from treated and control units are driven by some common factors and there they more likely to move ups/downs together. Imposing correction restriction can improve out-of-sample forecast. Therefore, in this section we compare the out-of-sample forecast performances of the modified synthetic control method and the least squares method. Specifically, we choose a value $T_0 \in (1, T_1) = (1, 90)$ to estimate the regression model, then we forecast outcome y_{1t} for $t = T_0 + 1, ..., T_1$. Since there is not treatment prior to T_1, we can compare the average prediction

squared error over the period $t = T_0 + 1, ..., T_1$. Specifically, we estimate the following model

$$y_t = x_t'\beta + u_{1t} \qquad\qquad t = 1, ..., T_0 \tag{6.2}$$

by the modified synthetic control and the least squares method. Let $\hat{\beta}_{T_0}$ and $\hat{\beta}_{OLS}$ denote the resulting estimators using the two methods, respectively. We prediction y_{1t}^0 by $\hat{y}_{1t,Msyn}^0 = x_t'\hat{\beta}_{T_0}$ and $\hat{y}_{1t,OLS}^0 = x_t'\hat{\beta}_{OLS}$ for $t = T_0 + 1, ..., T_1$. Then we computet the prediction MSEs by

$$\begin{aligned} PMSE_{Msyn} &= \frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,Msyn}^0)^2, \\ PMSE_{OLS} &= \frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,OLS}^0)^2. \end{aligned}$$

As suggested in Li and Bell (2017), we consider the cases that the 'pre-treatment' estimation sample is larger than the 'post-treatment' evaluation sample. We choose six different values for $T_0 = \{60, 65, 70, 75, 89, 85\}$. The corresponding evaluation sample sizes are $T_1 - T_0 = \{30, 25, 20, 15, 10, 5\}$. We report the ratio of PMSE as $PMSE_{OLS}/PMSE_{Msyn}$. The results are reported in Table 6.

Table 6: Out-of-sample Prediction MSE ratio

T_0	60	65	70	75	80	85
$\frac{PMSE_{OLS}}{PMSE_{Msyn}}$	1.680	1.104	1.020	1.273	1.188	1.143

From Table 6 we observe that the least squares method has larger PMSE than the modified synthetic control method for all cases. The PMSE for the former ranges from 2% to 68% larger than the later. Thus, the empirical example shows that, in order to more accurately predict the counterfactual outcomes for the treated unit, it is helpful to impose non-negativity restriction on the slope coefficients when estimating model (6.2).

7 Conclusion

The synthetic control method is a popular and powerful way for estimating average treatment effects. In this paper we contribute to the theoretical analysis of the synthetic control method. Under the assumption that there is a long panel data with large pre and post-treatment periods, using the projection theory, we derive the asymptotic distribution of the synthetic control ATE estimator. The asymptotic distribution is non-normal and non-standard. Moreover, it is known that the standard bootstrap method does not work in this case. We resolve the difficulty by proposing an easy-to-implement sub-sampling method and we establish the validity of sub-sampling method in inferences. When one only has a long pre-treatment data, but a short post-treatment data, as suggested by Ferman and Pinto (2016a), Andrews' (2003) end of sample instability test can be used to test the null

hypothesis for a given constant level of treatment effect ($\Delta_{1t} = \Delta_{1,0}$ for all $t = T_1 + 1, ..., T$).

We also prove that, when the pre-treatment sample size is large than the number of control units (i.e., $T_1 > N - 1$), the synthetic control estimator, as a constrained minimization problem, has an unique solution under a mild condition that the T_1 by N data matrix has a full column rank. We further show the modified synthetic control method suggested by Doudchenko and Imbens (2016) can give reliable ATE estimation results even when the "parallel lines" assumption is violated for the standard synthetic control method. Simulations show that the modified synthetic control method performs well in practice. Finally, we apply the synthetic control method to a marketing data estimating ATE of opening a showroom at Columbus by WarbyParker.com. The empirical application demonstrates that when the standard synthetic control method fits the data poorly, the modified synthetic control method fits the data well and gives reasonable ATE estimation results.

Appendix A: Uniqueness of the synthetic control estimator

A.1 Assumptions

We first list assumptions that are used in deriving the main results of the paper.

Assumption 1. The data $\{x_t\}_{t=1}^T$ is a weakly dependent stationary process so that laws of large number holds: $T_1^{-1}\sum_{t=1}^{T_1} x_t \xrightarrow{p} E(x_t)$ and $(X'X/T_1) \equiv T_1^{-1}\sum_{t=1}^{T_1} x_t x_t' \xrightarrow{p} E(x_t x_t')$, $E(x_t x_t')$ is positive definite, where X is the $T_1 \times N$ matrix with its t^{th} row given by $x_t' = (1, y_{2t}, ..., y_{Nt})$. Let $\eta = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, then η is a finite non-negative constant.

Assumption 2. $\{u_{1t}\}_{t=1}^T$ is zero mean and serially uncorrelated satisfying $T_1^{-1/2}\sum_{t=1}^{T_1} x_t u_{1t} \xrightarrow{d} N(0, \Sigma)$, where $\Sigma = \lim_{T_1\to\infty} T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1} E(u_{1t}u_{1s}x_t x_s')$.

Assumption 3. Let $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$, we assume that v_{1t} has zero mean and satisfies a central limit theorem: $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Sigma_2)$, where $\Sigma_2 = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$.

Assumption 4. Let $w_t = (y_{1t}, y_{2t}, ..., y_{Nt}, \Delta_{1t}d_t)$ for $t = 1, ..., T$, where $d_t = 0$ if $t \le T_1$ and $d_t = 1$ if $t \ge T_1 + 1$. Assume that $\{w_t\}_{t=1}^{T_1}$ and $\{w_t\}_{t=T_1+1}^{T}$ are both weakly stationary processes. Define $\rho(\tau) = \max_{1\le t\le T}\max_{1\le i,j\le N+1}|Cov(w_{it}, w_{j,t+\tau})/\sqrt{Var(w_t)Var(w_{j,t+\tau})}|$. Then there exits some finite positive constants $C > 0$, $0 < \lambda < 1$ such that $\rho(\tau) < C\lambda^\tau$.

Assumptions 1 and 2 imply that $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} N(0, A^{-1}\Sigma A^{-1})$, where $A = E(x_t x_t')$ and Σ is defined in assumption 2. Assumption 3 requires that a central limit theorem applies to a partial sum of v_{1t}. Assumption 4 is also used in Li and Bell (2017), this assumption ensures that the estimator $\hat{\beta}_{T_1}$ using the pre-treatment data is asymptotically independent with an quantity that involves the post-treatment sample average of the de-mean treatment effects and the idiosyncratic error.

A.2 A projection of the unconstrained estimator

We write the regression model in a matrix form:

$$Y = X\beta_0 + u,$$

where Y and u are both $T_1 \times 1$ vectors, X is of dimension $T_1 \times N$ and has a full column rank, β_0 is of dimension $N \times 1$. It is assumed that the true parameter $\beta_0 \in \Lambda$, where Λ is a closed and convex set ($\Lambda = \Lambda_{Syn}$ or Λ_{Msyn} in our applications).

We denote the constrained least squares estimator as $\hat{\beta}_{T_1}$, i.e.,

$$\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(Y - X\beta)^{'}(Y - X\beta) \equiv \arg\min_{\beta\in\Lambda}\|Y - X\beta\|^2,$$

where $||A||^2 = A^{'}A$ for a vector A.

We denote the unconstrained least squares estimator as $\hat{\beta}_{OLS} = arg\min_{\beta\in\mathcal{R}^N}(Y - X\beta)^{'}(Y - X\beta)$, i.e., $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. By the definition of $\hat{\beta}_{OLS}$, we may write

$$Y = X\hat{\beta}_{OLS} + \hat{u},$$

where $\hat{u} = Y - X\hat{\beta}_{OLS}$. It follows that

$$\begin{aligned} f(\beta) &\overset{def}{=} \|Y - X\beta\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta) + \hat{u}\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta)\|^2 + 2\hat{u}^{'}X(\hat{\beta}_{OLS} - \beta) + \|\hat{u}\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta)\|^2 + \|\hat{u}\|^2 \\ &\equiv (\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta) + \|\hat{u}\|^2, \end{aligned} \tag{A.1}$$

where the fourth equality follows from $\hat{u}^{'}X = 0$ (least squares residual $\hat{u}$ is orthogonal to X).

Since $\|\hat{u}\|^2$ is unrelated to β, the minimizer of $f(\beta)$ is identical to the minimizer of $(\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta)$. Thus, we have

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS} - \beta)^{'}(X^{'}X/T_1)(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta\in\Lambda}||\hat{\beta}_{OLS} - \beta||_X^2, \end{aligned}$$

where the second equality follows since $T_1 > 0$.

A.3 The uniqueness of the (modified) synthetic control estimator

We first give the definition of a strictly convex function. A function f is said to be *strictly convex* if $f(\alpha x + (1-\alpha)y)) < \alpha f(x) + (1-\alpha)f(y)$ for all $0 < \alpha < 1$ and for all $x \neq y$, $x, y \in D$, where D is the domain of f.

Under the assumption that the data matrix $X_{T_1 \times N}$ has a full column rank, we show below that $f(\beta) \stackrel{def}{=} \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ is a strictly convex function. Since the objective function is a convex function and the constrained domain for β, Λ_{Syn} and Λ are convex sets, then the constrained minimization problem has an unique (global) minimizer. To see this, we argue by contradiction. Suppose that we have two local minimizers $z_1 \neq z_2$. Then for any convex combination $z_3 = \alpha z_1 + (1-\alpha)z_2$, we have $f(z_3) < \alpha f(z_1) + (1-\alpha)f(z_2)$ for all $\alpha \in (0,1)$. This contradicts the fact that z_1 and z_2 are two minimizers. Hence, we must have $z_1 = z_2$ and the minimizer is unique.

It remains to show that $f(\beta)$ defined in (A.1) is a strictly convex function. To see this, letting $A = X'X$, then for β, $\gamma \in \mathcal{R}^N$ with $\beta \neq \gamma$ and all $\alpha \in (0,1)$, we have (we ignore the irrelevant constant term $\|\hat{u}\|^2$ in $f(\beta)$)

$$\begin{aligned}
f(\alpha\beta + (1-\alpha)\gamma) &= (\alpha(\hat{\beta}_{OLS} - \beta) + (1-\alpha)(\hat{\beta}_{OLS} - \gamma))' A(\alpha(\hat{\beta}_{OLS} - \beta) + (1-\alpha)(\hat{\beta}_{OLS} - \gamma)) \\
&= \alpha^2(\hat{\beta}_{OLS} - \beta)' A(\hat{\beta}_{OLS} - \beta) + (1-\alpha)^2(\hat{\beta}_{OLS} - \gamma)' A(\hat{\beta}_{OLS} - \gamma) \\
&\quad +2\alpha(1-\alpha)(\hat{\beta}_{OLS} - \beta)' A(\hat{\beta}_{OLS} - \gamma) \\
&< \alpha^2 f(\beta) + (1-\alpha)^2 f(\gamma) + \alpha(1-\alpha)[f(\beta) + f(\gamma)] \\
&= \alpha f(\beta) + (1-\alpha) f(\gamma),
\end{aligned}$$

where the above inequality follows from that A is positive definite and $\beta \neq \gamma$. Hence,

$$\begin{aligned}
0 &< (\beta - \gamma)' A(\beta - \gamma) \\
&= ((\beta - \hat{\beta}_{OLS}) - (\gamma - \hat{\beta}_{OLS}))' A((\beta - \hat{\beta}_{OLS}) - (\gamma - \hat{\beta}_{OLS})) \\
&= (\beta - \hat{\beta}_{OLS})' A(\beta - \hat{\beta}_{OLS}) + (\gamma - \hat{\beta}_{OLS})' A(\gamma - \hat{\beta}_{OLS}) \\
&\quad -2(\beta - \hat{\beta}_{OLS})' A(\gamma - \hat{\beta}_{OLS}) \\
&= f(\beta) + f(\gamma) - 2(\hat{\beta}_{OLS} - \beta)' A(\hat{\beta}_{OLS} - \gamma).
\end{aligned}$$

This proves that $f(\cdot)$ is a strictly convex function.

Appendix B: Proofs of Theorems 3.2, 3.3 and 4.1

B.1 Proof of Theorem 3.2

Continuing with the setup in Appendix A, the constrained estimator is defined by

$$\hat{\beta}_{T_1} = \arg\min_{\beta \in \Lambda} (\beta - \hat{\beta}_{OLS})' (X'X/T_1)(\beta - \hat{\beta}_{OLS}). \tag{B.1}$$

Thus, $\hat{\beta}_{T_1}$ is the projection onto Λ with respect to the norm $\|a\| = \sqrt{a'(X'X/T_1)a}$ which is random, rendering the theory in Fang and Santos (2015) not directly applicable. However, since $X'X/T_1 \xrightarrow{p} E(X_tX_t')$, we will show that one can replace $X'X/T_1$ by $E(X_tX_t')$ without affecting the asymptotic results. Define the following "infeasible estimator" (it is infeasible because $E(X_t'X_t)$ is unknown in practice):

$$\tilde{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(\beta - \hat{\beta}_{OLS})'E(X_tX_t')(\beta - \hat{\beta}_{OLS}) = \Pi_\Lambda \hat{\beta}_{OLS} \ , \tag{B.2}$$

where Π_Λ is the projection onto Λ with respect to the norm $\|a\| = \sqrt{a'E(X_tX_t')a}$, i.e.,

$$\Pi_\Lambda \beta = \arg\min_{\lambda\in\Lambda}(\beta - \lambda)'E(X_tX_t')(\beta - \lambda) \ . \tag{B.3}$$

By Lemma 4.6 of Zarantonello (1971, page 300), see also Proposition 4.1 of Fang and Santos (2015), we know that

$$\begin{aligned}
\sqrt{T_1}(\tilde{\beta}_{T_1} - \beta_0) &= \sqrt{T_1}(\Pi_\Lambda \hat{\beta}_{OLS} - \Pi_\Lambda \beta_0) \\
&= \sqrt{T_1}\Pi_{T_{\Lambda,\beta_0}}(\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\
&= \Pi_{T_{\Lambda,\beta_0}}\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\
&\xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1,
\end{aligned} \tag{B.4}$$

where the first equality follows from $\tilde{\beta}_{T_1} = \Pi_\Lambda \hat{\beta}_{OLS}$ and $\beta_0 \in \Lambda$ so that $\beta_0 = \Pi_\Lambda \beta_0$.

In Lemma C.1 of the supplementary Appendix C we show that

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}). \tag{B.5}$$

Theorem 3.2 follows from (B.4) and (B.5).

We give some explanations of the above derivations. Hilbert Space projection onto convex sets was studied by Zarantonello (1971) and extended to general econometric modeling settings by Fang and Santos (2015). The projection operator Π_Λ: $\mathcal{R}^N \to \Lambda$ (Λ is convex subset in $\mathcal{R}^N$) can be viewed as a functional mapping. Zarantonello (1971) showed that Π_Λ is (Hadamard) directional differentiable, and its directional derivative at $\beta_0 \in \Lambda$ is $\Pi_{T_{\Lambda,\beta_0}}$, the projection onto the tangent cone of Λ at β_0. Hence, the second equality of (B.4) follows from a functional Taylor expansion, the third equality follows from that T_{Λ,β_0} is positive homogenous of degree one.[3] The last line follows from $\sqrt{T_1}(\hat{\beta}_{OLS} - \Lambda\beta_0) \xrightarrow{d} Z_1$ and the continuous mapping theorem because projection is a continuous mapping.

We can also see the term 'tangent cone' is similar to what we term the derivative of a function at a given point as a 'tangent line' of the function at the given point. Now, the functional derivative of the mapping Π_Λ is a projection onto the cone $\Pi_{T_{\Lambda,\beta_0}}$ (rather than a line). Therefore, it is called the 'tangent cone' of Λ at β_0 and is denoted as T_{Λ,β_0}.

[3]The Projection T_{Λ,β_0} is not a linear operator. However, for $\alpha \geq 0$, we have $\alpha\, T_{\Lambda,\beta_0}\, \theta = T_{\Lambda,\beta_0}\, \alpha\, \theta$ for all $\theta \in \mathcal{R}^N$.

For readers' convenience, we give the formal definition of tangent cone of Λ at $\theta \in \mathcal{R}^N$ below:

$$T_{\Lambda,\theta} = \overline{\cup_{\alpha \geq 0} \alpha\{\Lambda - \Pi_\Lambda \theta\}}, \tag{B.6}$$

where for any set $A \in \mathcal{R}^N$, $\overline{A}$ is the closure of A.

It can be easily checked that for our synthetic control estimation problem, the tangent cone of Λ at β_0 is the same as the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

B.2 Proof of Theorem 3.3

First, we write $\hat{A} = \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ defined in (4.2) as $\hat{A} = \hat{A}_1 + \hat{A}_2$, where

$$\begin{aligned} \hat{A}_1 &= -\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{\frac{T_2}{T_1}}\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \\ \hat{A}_2 &= \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}. \end{aligned} \tag{B.7}$$

We know that $\hat{A}_2 \xrightarrow{d} Z_2$ by assumption 2, where Z_2 is distributed as $N(0, \Sigma_2)$ with $\Sigma_2 = E(v_{1t}^2)$.

By Theorem 3.2 and assumption 1, we have $\hat{A}_1 \xrightarrow{d} A_1 = -\eta E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1$, where $\eta = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$ and Z_1 is the weak limit of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$.

Also, by Lemma A.1 and Theorem 3.2 of Li and Bell (2016), we know that Z_1 and Z_2 are asymptotically independent with each other, this implies that $A_1 = -\eta E(x_t)\Pi_{T_{\Lambda,\beta_0}} Z_1$ is asymptotically independent of Z_2. Hence, we have

$$\hat{A} \xrightarrow{d} -\eta E(x_t)\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2. \tag{B.8}$$

B.3 Proof of Theorem 4.1

The proof that $\hat{A}^*$ can be used to approximate the distribution of $\hat{A}$ consists of the following arguments. First, we show that one can consistently estimate Σ_2 by $\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{v}_{1t}^2$, where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. From $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0 = x_t'(\beta_0 - \hat{\beta}_{T_1}) + \Delta_{1t} + u_{1t} = \Delta_{1t} + u_{1t} + O_p(T_1^{-1/2})$ and $\bar{\hat{\Delta}}_1 = \bar{x}'(\beta_0 - \hat{\beta}_{T_1}) + \bar{\Delta}_1 + \bar{u}_1 = \Delta_1 + O_p(T_1^{-1/2} + T_2^{-1/2})$, we have

$$\begin{aligned} \hat{\Sigma}_2 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\hat{\Delta}_{1t} - \hat{\Delta}_1)^2 \\ &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\Delta_{1t} + u_{1t} - \Delta_1)^2 + O_p(T_1^{-1/2} + T_2^{-1/2}) \\ &= \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t}^2 + O_p(T_1^{-1/2} + T_2^{-1/2}) \\ &= \Sigma_v + O_p(T_1^{-1/2} + T_2^{-1/2}). \end{aligned} \tag{B.9}$$

Next, it is obvious that $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t}^* \overset{d}{\sim} T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \overset{d}{\to} Z_2$, where $A \overset{d}{\sim} B$ means that A and B have the same asymptotic distribution. Hence, to show that the limiting distribution of $\hat{A}^*$ defined in (4.3) is the same as that of $\hat{A}$, we only need to show that $\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) \overset{d}{\sim} \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$. To see that this is indeed true, we have

$$\begin{aligned}
\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) &= \sqrt{m}(\hat{\beta}_m^* - \beta_0) + \sqrt{m}(\beta_0 - \hat{\beta}_{T_1}) \\
&= \sqrt{m}(\hat{\beta}_m^* - \beta_0) + O_p((m/T_1)^{1/2}) \\
&\overset{d}{\sim} \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \\
&\overset{d}{\to} \Pi_{T_\Lambda, \beta_0} Z_1, \qquad \text{(B.10)}
\end{aligned}$$

where the asymptotic equivalence $\overset{d}{\sim}$ follows the facts that $\sqrt{m}(\hat{\beta}_m - \beta_0)$ and $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ have the same asymptotic distribution, $\hat{\beta}_{T_1} - \beta_0 = O_p(T_1^{-1/2})$ and $m/T_1 = o(1)$; the last convergence result follow from Theorem 3.2.

It follows from (B.10) that $\hat{A}^*$ defined in (4.3) and $\hat{A}$ defined in (4.2) have the same asymptotic distribution. This completes the proof of Theorem 4.1.

8 References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72, 1-19.

Abadie, A., A. Diamond and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105, 493-505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review* 93, 113-132.

Andrews, D.W.K. 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter Space. *Econometrica* 68, 399-405.

Andrews, D.W.K. 2003. End-of-sample instability tests. *Econometrica* 71, 16611694.

Ashenfelter O (1978) Estimating the Effects of Training Programs on Earnings. *The Review of Economics and Statistics*. 60, 47-57.

Ashenfelter O, Card D (1985) Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. 1985. *The Review of Economics and Statistics*. 67, 648-660.

Athey, S., and G. Imbens (2016). The State of Applied Econometrics: Causality and Policy Evaluation. Working Paper.

Bickel, J.P. and A. Sakov. 2008. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18, 967-985.

Bradley, E., Hastie, T., Johnstone, T., and R. Tibshirani (2004). Least Angle Regression. *Annals of Statistics* 32, 407-499.

Busse, M., Silva-Risso J, Zettlemeyer F (2006) $ 1,000 Cash Back: The Pass-Through of Auto Manufacturer Promotions. *The American Economic Review*. 96, 1253-1270.

Chevalier, J, Mayzlin D (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*. 48, 345-354.

Conley, T.G. and C.R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93, 113-125.

Doudchenko, N. and G. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.

Du, Z. and L. Zhang (2015). Home-Purchase Restriction, Property Tax and Housing Price in China: A Counterfactual Analysis. *Journal of Econometrics*. 188, 558-568

Fang, Z. and A. Santos (2015). Inference on directionally differentiable functions. Under revision for *Review of Economic Studies*.

Ferman, B. and C. Pinto (2016a). Inference in Differences-in-difference with Few Treated Groups and Heteroskedasticity. Unpublished working paper.

Ferman, B. and C. Pinto (2016b). Revisiting the Synthetic Control Estimator. Unpublished working paper.

Forman C, Ghose A, Goldfarb A (2009) Competition Between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live. *Management Science* 55, 47-57.

Goldfarb A, Tucker C (2011) Private Regulation and Online Advertising. *Management Science*. 57, 40-56.

Hahn, J. and R. Shi (2016). Synthetic Control and Inference. Unpublished working paper.

Hsiao, C., Ching, H.S. and S.K. Wan (2012). A panel data approach for program evaluation: Measuring the benefit of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics* 27, 705-740.

Hong, H. and J. Li 2017. The Numerical Delta Method and Bootstrap Unpublished working paper.

Li, K. and D. Bell. 2017. Estimation of average treatment effects with panel data: asymptotic theory and implementation. *Journal of Econometrics* 197, 65-75.

Mantin, B. and E. Rubin. 2016. Fare Prediction Websites and Transaction Prices: Empirical Evidence from the Airline Industry *Marketing Sciences* 35, 640-655.

Ozturk, O.C., S. Venkataraman and P. K. Chintagunta. 2016. Price Reactions to Rivals Local Channel Exits *Marketing Sciences* 35, 588604.

Politis, D.N., J.P. Romano and M. Wolf. 1999. Sub-sampling. Springer.

Wang, K. and A. Goldfarb. 2017. Can offline stores drive online sales? forthcoming in *Journal of Marketing Research*

Zarantonello, E.H. (1971). Projection on convex sets and Hilbert Spaces and spectral theory. In *Contributions to Nonlinear Functional Analysis* (E.H. Zarantonello ed.). Academic Press.

Supplementary Appendix C: Two useful lemmas

In this supplementary appendix we prove two lemmas that are used to prove Theorem 3.2.

Lemma C.1 *Under the same conditions as in Theorem 3.2, we have*

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}).$$

Proof: For any fixed $\epsilon > 0$, suppose that $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$, then we have

$$\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) < \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \ , \quad \text{(C.1)}$$

where the strict inequality is due to uniqueness of the projection and the assumption that $\epsilon > 0$ which implies that $\hat{\beta}_{T_1} \neq \tilde{\beta}_{T_1}$. By simple algebra (adding/subtracting terms), we have:

$$\begin{aligned}
& \sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \ & \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \ & \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
& + \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}). \qquad \text{(C.2)}
\end{aligned}$$

By (C.1) and (C.2) we know that the sum of the last two terms in (C.2) is negative, i.e.,

$$\begin{aligned}
D_{T_1} \ & \stackrel{def}{=} \ \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\
& \equiv \ D_{1,T_1} + D_{2,T_1} < 0. \qquad \text{(C.3)}
\end{aligned}$$

Let $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$, the unit sphere in $\mathcal{R}^N$, we have:

$$\begin{aligned}
D_{1,T_1} \ & = \ \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& = \ \|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|^2 \left[\frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|} \left(\frac{1}{T_1}X'X\right) \frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|}\right] \\
& \geq \ T_1\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\|^2 \inf_{a \in \mathcal{S}^N} a' \left(\frac{1}{T_1}X'X\right) a \\
& = \ T_1\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\|^2 \lambda_{\min}\left(\frac{1}{T_1}X'X\right) \\
& \geq \ \epsilon^2 \lambda_{\min}\left(\frac{1}{T_1}X'X\right) \\
& \xrightarrow{p} \ \epsilon^2 \lambda_{\min}[E(X_t X_t')] > 0 \ , \qquad \text{(C.4)}
\end{aligned}$$

because $\sqrt{T_1}\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\| \geq \epsilon$ and $E(X_t X_t')$ is nonsingular, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a square matrix A, the third equality used Lemma C.2 which is proved at the end of this Appendix.

By writing $(X'X/T_1) = E(X_t X_t') + (X'X/T_1) - E(X_t X_t')$, the second term in (C.3) can be rewritten as:

$$\begin{aligned} D_{2,T_1} &= 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &= 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_t X_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &\quad +2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1 - E[X_t X_t'])\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &= 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_t X_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) + o_p(1) \ , \end{aligned} \tag{C.5}$$

where we used the fact that $\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) = O_p(1)$ because:

$$\begin{aligned} \|\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1})\| &\leq \|\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)\| + \|\sqrt{T_1}(\tilde{\beta}_{T_1} - \beta_0)\| \\ &= \|\sqrt{T_1}(\Pi_{\Lambda,T_1}\hat{\beta}_{OLS} - \beta_0)\| + \|\sqrt{T_1}(\Pi_{\Lambda}\hat{\beta}_{OLS} - \beta_0)\| \\ &\leq \sqrt{T_1}\|\hat{\beta}_{OLS} - \beta_0\|_{T_1} + \sqrt{T_1}\|\hat{\beta}_{OLS} - \beta_0\| = O_p(1) \ , \end{aligned}$$

where we used the Lipschitz continuity of projection operators (Zarantonello, 1971; p.241, first display in equation (1.8)), and Π_{Λ,T_1} is the projection onto Λ with respect to the aforementioned random norm $\|a\|_{T_1} = \sqrt{a'(X'X/T_1)a}$.

Also, by the definition of $\tilde{\beta}_{T_1}$ and Lemma 1.1 in Zarantonello (1971, page 239),

$$\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_t X_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \geq 0 \ . \tag{C.6}$$

Combining (C.3), (C.4), (C.5) and (C.6), we know that

$$\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_t X_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) = o_p(1). \tag{C.7}$$

Thus, we have shown that: if $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$, then

$$\begin{aligned} &\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{\text{OLS}})'\left(\frac{1}{T_1}X'X\right)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) - \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{\text{OLS}})'\left(\frac{1}{T_1}X'X\right)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\ = \ & D_{T_1} < 0, \end{aligned} \tag{C.8}$$

which implies that (if A implies B, then $P(A) \leq P(B)$)

$$\begin{aligned} P(\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon) &\leq P(D_{T_1} < 0) \\ &\leq P(o_p(1) + \epsilon^2\lambda_{\min}\left(\frac{1}{T_1}X'X\right) < 0) \\ &\to P(\epsilon^2\lambda_{\min}\left(E(X_t X_t')\right) \leq 0) = 0. \end{aligned} \tag{C.9}$$

Equation (C.9) is equivalent to $\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} = o_p(T_1^{-1/2})$, or

$$\hat{\beta}_{T_1} = \Pi_{\Lambda,T_1}\hat{\beta}_{OLS} + o_p(T_1^{-1/2}) \ . \tag{C.10}$$

This finishes the proof of Theorem 3.2.

Lemma C.2 *Let A be an $N \times N$ positive definite matrix, $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$, the unit sphere in $\mathcal{R}^N$, then we have* $\inf_{a\in\mathcal{S}^N} a'Aa = \lambda_{min}(A)$.

Proof: Let $v_1, ..., v_N$ be N eigen-vectors of A with corresponding eigen-values $\lambda_1, ..., \lambda_N$ so that $Av_j = \lambda_j v_j$ for $j = 1, ..., N$. Then since $v_1, ..., v_N$ form an orthonormal base for $\mathcal{S}^N$, we have for any $a \in \mathcal{S}^N$, $a = \sum_{i=1}^N c_i v_i$ with $\sum_{i=1}^N c_i^2 = 1$ since $a'a = 1$ and $v_i'v_j = \delta_{ij}$ (the Kronecker delta). Then we have

$$\begin{aligned} a'Aa &= \sum_{i=1}^N \sum_{j=1}^N c_i v_i' A c_j v_j \\ &= \sum_{i=1}^N \sum_{j=1}^N c_i v_i' c_j A v_j \\ &= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \lambda_j v_i' v_j \\ &= \sum_{i=1}^N \lambda_j c_j^2 \\ &\geq \lambda_{min} \sum_{j=1}^N c_j^2 \\ &= \lambda_{min}, \end{aligned} \tag{C.11}$$

which implies (i) $\inf_{a\in\mathcal{S}^N} a'Aa \geq \lambda_{min}$.

On the other hand, pre-multiplying $Av_j = \lambda_j v_j$ by v_j', we get $\lambda_j = v_j'Av_j \geq \inf_{a\in\mathcal{S}^N} a'Aa$ for all $j = 1, ..., N$, which implies (ii) $\lambda_{min} \geq \inf_{a\in\mathcal{S}^N} a'Aa$. Combining (i) and (ii) we finish the proof of Lemma C.2.

Supplementary Appendix D: Asymptotic theory with trend stationary data

Asymptotic theory for the unconstrained estimator

In this section, we derive the asymptotic distribution of the ATE estimator $\hat{\Delta}_1$ defined in (3.9). For the post-treatment, we have $y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$. Hence, we have for $t = 1, ..., T$,

$$y_{1t} = \alpha t + z_t'\beta + d_t\Delta_{1t} + v_{1t}, \tag{D.1}$$

where $d_t = 0$ for $t \leq T_1$ and $d_t = 1$ for $t \geq T_1 + 1$.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β based on (3.8). Then it is well established that (e.g., Hamilton (1994), Chapter 16) $\hat{\alpha} - \alpha = O_p(T_1^{-3/2})$ and $\hat{\beta} - \beta = O_p(T_1^{-1/2})$. Thus, using (3.9) and (D.1) we have

$$\begin{aligned}
\hat{\Delta}_1 - \Delta_1 &= \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left[y_{1t} - \hat{y}_{1t}^0 \right] - \Delta_1 \\
&= -\frac{1}{T_2} \sum_{t=T_1+1}^{T} \left[(\hat{\alpha}_{T_1} - \alpha_0) t - z_t' (\hat{\beta}_{T_1} - \beta_0) + \Delta_{1t} - \Delta_1 + v_{1t} \right] \\
&= -\left[\frac{2T_1 + T_2 + 1}{2} \right] (\hat{\alpha} - \alpha) - [E(z_t') + o_p(1)](\hat{\beta} - \beta) + \frac{1}{T_2} \sum_{t=T_1+1}^{T} v_{1t}, \qquad \text{(D.2)}
\end{aligned}$$

where we used $\sum_{t=T_1+1}^{T} t = (T_1 + 1 + T)T_2/2 = (2T_1 + T_2 + 1)T_2/2$, $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$.

Hence,

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -\sqrt{T_2/T_1} \left[\frac{2 + T_2/T_1}{2} \right] \sqrt{T_1^3}(\hat{\alpha}_{T_1} - \alpha_0) - \sqrt{T_2/T_1} E(z_t') \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \\
&\quad + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -\left(\sqrt{T_2/T_1}(2 + T_2/T_1)/2 \,,\ \sqrt{T_2/T_1} E(z_t') \right) \begin{pmatrix} \sqrt{T_1^3}(\hat{\alpha}_{T_1} - \alpha_0) \\ \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \end{pmatrix} \\
&\quad + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -c' M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1), \qquad \text{(D.3)}
\end{aligned}$$

where $c = (\sqrt{\eta}(2 + \eta)/2, \sqrt{\eta} E(z_t'))'$, $\eta = \lim_{T_1, T_2 \to \infty} T_2/T_1$, $\hat{\gamma}_{T_1} = (\hat{\alpha}_{T_1}, \hat{\beta}_{T_1}')'$ and $\gamma_0 = (\alpha_0, \beta_0')'$, $M_{T_1} = \sqrt{T_1} diag(T_1, 1, ..., 1)$ which is an $(N + 1) \times (N + 1)$ diagonal matrix with the first diagonal element equals to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$.

To establish the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we make the following assumptions.

Assumption D1. Let $z_t = (1, y_{2t}^{*\prime}, ..., y_{Nt}^*)'$. We assume that (i) $\{z_t\}_{t=1}^{T}$ is a weakly dependent and weakly stationary process, $T_1^{-1} \sum_{t=1}^{T_1} z_t z_t' \xrightarrow{p} E(z_t z_t')$ as $T_1 \to \infty$, and $[E(z_t z_t')]$ is invertible; (ii) $M_{T_1}(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{d} N(0, \Omega)$, where Ω is a positive definite matrix.

Assumption D2. Let $v_{1t} = \Delta_{1t} - \Delta_1 + v_{1t}$. Then $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Sigma_2)$ as $T_2 \to \infty$, where $\Sigma_2 = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(v_{1t} v_{1s})$ is the asymptotic variance of $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}$.

Assumption D3. Let $w_t = (v_{1t}, y_{2t}^*, ..., y_{Nt}^*)'$. We assume that w_t is a ρ-mixing process with the mixing coefficient $\rho(\tau)$ satisfies the condition: $\rho(\tau) \leq C \lambda^{\tau}$ for some finite positive constants $C > 0$ and $0 < \lambda < 1$, where $\rho(\tau) = \max_{1 \leq i,j \leq N} |Cov(w_{it}, w_{j,t+\tau})| / \sqrt{Var(w_{it}) Var(w_{j,t+\tau})}$, w_{it} is the i^{th}

component of w_t for $i = 1, ..., N$.

Assumptions D1 and D2 are not restrictive. They require that (z_t, v_{1t}) to be a weakly dependent stationary process so that law of large numbers and central limit theorem hold for their (partial) sums. If $E(z_t z_t')$ is not invertible, we can remove the linearly dependent regressors and redefine z_t as a subset of $(1, y_{2t}^*, ..., y_{Nt}^*)'$ such that assumption 1 holds. Assumption D3 further imposes an exponential decay rate for the ρ-mixing processes. Many ARMA processes are known to be ρ-mixing with exponential decay rate.

By Assumption D3 and the proof of Theorem 3.2 and Lemma 1 in Li and Bell (2017), we know that $\hat{\gamma} - \gamma$ is asymptotic independent with $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}$. Therefore, applying the projection theory to (D.3) we immediately have the following result.

Under assumptions D1 to D3 and note that $\gamma_0 \in \Lambda$, we have

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -c' M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -c' M_{T_1}(\Pi_\Lambda \hat{\gamma}_{OLS} - \Pi_\Lambda \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -c' \Pi_{T_{\Lambda,\gamma_0}} M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&\xrightarrow{d} -c' \Pi_{T_{\Lambda,\gamma_0}} Z_3 + Z_2, \qquad \text{(D.4)}
\end{aligned}$$

where Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in Assumption C1, Z_2 is independent with Z_3, and is normally distributed with a zero mean and variance $\Sigma_2 = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(v_{1t} v_{1s})$.

Supplementary Appendix E: Explanation of sub-sampling method works for a wide range of sub-sample sizes

In this appendix, we explain why the sub-sampling method works well for our estimated ATE estimator for a wide range of sub-sample size m values.

E.1 A simple example from Andrews (2000)

We consider a simple example as considered in Andrews (2000) where Y_i, for $i = 1, ..., n$, is iid $N(\mu_0, 1)$ with $\mu_0 \geq 0$, i.e., $Y_i = \mu_0 + u_i$ with u_i iid $N(0, 1)$ and that $\mu_0 \in \Lambda = \mathcal{R}^+ \stackrel{def}{=} \{y : y \geq 0\}$. The constrained least squares estimator of μ_0 is $\hat{\mu}_n = \max\{\bar{Y}_n, 0\}$, where $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$. It is easy to show that

$$\hat{S}_n \stackrel{def}{=} \sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{d} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases} \qquad \text{(E.1)}$$

where Z denotes a standard normal random variable. Let Y_i^* be random draws from $\{Y_j\}_{j=1}^n$, then a bootstrap analogue of $(E.1)$ is $\sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$, where $\hat{\mu}_n^* = \max\{\bar{Y}_n^*, 0\}$, where $\bar{Y}_n^* = n^{-1}\sum_{i=1}^n Y_i^*$. Andrews (2000) show that this standard resampling bootstrap method as well as several parametric bootstrap methods do not work in the sense that when $\mu_0 = 0$, $\tilde{S}_n^* = \sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$ will not converge to $\max\{Z, 0\}$, the limiting distribution of $\hat{S}_n$. In fact, Andrews shows that $\hat{S}_n^*$ converges to a distribution that is to the left of $\max\{Z, 0\}$.

Andrews (2000) also suggests a few re-sampling methods that overcome the problem. One particular easy-to-implement method is a parametric sub-sampling method. Specifically, for m satisfies that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, one can use $\tilde{S}_m^* = \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n)$ to approximate the distribution of $\sqrt{n}(\hat{\mu}_n - \mu_0)$, where $\hat{\mu}_m = \max\{\bar{Y}_m^*, 0\}$, $\bar{Y}_m^* = m^{-1}\sum_{i=1}^m Y_i^*$ with Y_i^* is iid draws from $N(\bar{Y}_n, 1)$, i.e., $Y_i^* = \bar{Y}_n + u_i^*$ with u_i^* iid $N(0,1)$. To see that the sub-sampling method indeed works, we have that, conditional on $\{Y_i\}_{i=1}^n$,

$$\begin{aligned}
\hat{S}_m^* &\overset{def}{=} \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n) \\
&= \max\{\sqrt{m}\,\bar{Y}_m^*, 0\} - \sqrt{m}\,\hat{\mu}_n \\
&= \max\{\sqrt{m}\,\bar{Y}_m^*, 0\} - \sqrt{m}\,\mu_0 - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n + \bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_0\} - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + \sqrt{m/n}\sqrt{n}(\bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_)\right\} - \sqrt{m/n}\sqrt{n}(\hat{\mu}_n - \mu_0) \\
&= \max\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + o_p(1), -\sqrt{m}\,\mu_0\} + o_p(1) \\
&\overset{d}{\to} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases}
\end{aligned} \tag{E.2}$$

where the second equality follows from the definition of $\hat{\mu}_m^*$, we add/subtract $\sqrt{m}\,\mu_0$ at the third equality, the fourth equality follows from $\max\{a, b\} - c = \max\{a - c, b - c\}$, the sixth equality follows from $m/n = o(1)$, $\sqrt{n}(\bar{Y}_n - \mu_0) = O_p(1)$ and $o(1)O_p(1) = o_p(1)$. The last equality follows from the fact that $Y_i^* - \bar{Y}_n = u_i^*$ is iid $N(0,1)$. Hence, $\sqrt{m}(\hat{Y}_m^* - \bar{Y}_n) \overset{d}{\sim} N(0,1) \equiv Z$ for any value of m. If $\{Y_i^*\}_{i=1}^m$ is iid with mean $\bar{Y}_n$ and unit variance but is not normally distributed, then we need m to be large so that $\sqrt{m}(\hat{Y}_m^* - \bar{Y}_n) \overset{d}{\to} N(0,1) \equiv Z$ by virtue of a central limit theorem argument (as $m \to \infty$).

Comparing (E.1) and (E.2), we see that sub-sampling method works under very mild conditions that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$.

E.2 Testing zero ATE by sub-sampling method

We conduct simulations to examine the finite sample performances of the sub-sampling method. We generate Y_i iid $N(0,1)$ (i.e., $\mu_0 = 0$) for $i = 1, ..., n$ and we choose $n = 100$ and conduct 5000 simulations. Within each simulation, we generate 2000 sub-sampling samples with sub-sample sizes $m \in \{5, 10, 20, 30, 50, 100\}$. Note that we select the largest $m = n = 100$ because we want to show

numerically that the standard bootstrap method does not work. For each fixed value of m, we sort the 2000 sub-sampling statistics in an ascending order such that $\hat{S}^*_{m,(1)} \leq \hat{S}^*_{m,(2)} \leq ... \leq \hat{S}^*_{m,(2000)}$, then we get right-tail α-percentile value by $\hat{S}^*_{((1-\alpha)(2000))}$. We record rejection rate as the percentage that $\hat{S}$ is greater or equal to $\hat{S}^*_{((1-\alpha)(2000))}$ for $\alpha \in \{0.01, 0.05, 0.1, 0.2\}$. We consider two cases: (i) We generate Y_i iid $N(0,1)$ and $Y_i^* = \bar{Y}_n + v_i$ with v_i iiid $N(0,1)$; and (ii) We generate Y_i uniformly distributed over $[-\sqrt{3}, \sqrt{3}]$ (so that it has zero mean and unit variance) and $Y_i^* = \bar{Y}_n + v_i$ with v_i iid uniformly distributed over $[-\sqrt{3}, \sqrt{3}]$. The results for the two cases are almost identical. To save space we only report are reported the normally distributed v_i in Table 7.

Table 7: Estimated sizes ($Y_i^* \sim N(\bar{Y}_n, 1)$)

	m=5	m=10	m=20	m=30	m=50	m=100
1%	.0132	. 0126	.0124	.0130	.0136	.0248
5%	.0516	.0518	.0518	.0532	.0658	.1032
10%	.0960	.0968	.1006	.1104	.1346	.2014
20%	.1936	.2004	.2278	.2588	.3164	.4020

First, we see that the sub-sampling method with $5 \leq m \leq 20$ seem to work well. Second, we see clearly that using $m = n$ or m close to n ($m \geq 50$) do not work. For example, when $m = n$, it gives estimated rejection rates double that of the nominal levels. Andrews (2000) showed that the distribution of $\sqrt{n}(\hat{\mu}^*_n - \hat{\mu}_n)$ is to the left of that of $\sqrt{n}(\hat{\mu}_n - \mu_0)$. Hence, the bootstrap method will lead to over rejection of the null hypothesis. Our simulation results verifies Andrews' theoretical analysis.

The simulation results seem to be in contradiction to the simulation results reported in Section 5 where even for $m = n$, the sub-sampling method seems to be fine. We explain the seemingly contradictory results in the next subsection.

E.3 Not all parameters are at the boundary

Our simulations reported in Section 5 corresponds to the case of $\beta_{0,j} > 0$ for $j = 2, ..., 7$ and $\beta_{0,j} = 0$ for $j = 8, ..., 11$. The constrained estimators $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 8, 9, 10, 11$ can cause problems for the standard bootstrap method not to work. However, notice that our ATE estimator also depends on $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 1, ..., 7$ which does not take the one-dimensional problem boundary value 0. This helps to improve sub-sampling method for large value of m. More importantly, our ATE estimator also contains a term not related to $\hat{\beta}_{T_1}$ (see the second term at the right hand side of (4.5)) and the existence of this term further improves the performance of the sub-sampling method when m is close to or equal to n. This is the reason why in our simulations even when $m = n$ the sub-sampling method seems to work fine. To numerically verify this conjecture, we generate a sequence of iid $Z_1, Z_2 \sim N(0, \sigma_v^2)$ random variables and add them to $\hat{S}_n$ and $\hat{S}^*_m$, i.e., $\tilde{S}_n = \hat{S}_n + Z_1$ and $\tilde{S}^*_m = \hat{S}^*_m + Z_2$, we then repeat the simulations to compute the estimated sizes. The results for $\sigma_v = 1$ and 5 are reported in Table 8. We observe the performance of the sub-sampling statistic $\tilde{S}^*_m$ improves significantly over $\hat{S}^*_m$ for

$m = 50$ and 100. Consider the case of $\sigma_v = 1$ and $m = n$, the rejection rates based on $\tilde{S}^*_m$ is about 20% higher than that of the nominal levels whereas it was 100% higher than that of nominal levels based on $\hat{S}^*_m$.

From Table 8 we see that when σ_v^2 is large, Z_1 and Z_2 becomes the dominating components of $\tilde{S}_n$ and $\tilde{S}^*_m$, therefore, the sub-sampling method works well for all values of m including $m = n$. Also note that the estimated sizes for $\sigma_v^2 = 1$ only slightly over sized compared to $\sigma_v^2 = 25$ shows that the significant improvements in the estimated sizes (over the case of $\sigma_v^2 = 0$) does not require adding a regular component with large dominating variance.

Table 8: Estimated sizes: Adding a $N(0, \sigma_v^2)$ to $\hat{S}_n$ and $\hat{S}^*_m$

	m=5	m=10	m=20	m=30	m=50	m=100
	$\sigma_v = 1$					
1%	.0104	.0110	.0112	.0128	.0122	.0114
5%	.0550	.0562	.0562	.0590	.0600	.0648
10%	.1066	.1098	.1140	.1168	.1198	.1236
20%	.2170	.2244	.2320	.2372	.2440	.2520
	$\sigma_v = 5$					
1%	.0112	.0116	.0116	.0110	.0124	.0128
5%	.0518	.0521	.0528	.0530	.0542	.0556
10%	.1030	.1044	.1046	.1048	.1060	.1074
20%	.2070	.2082	.2030	.2102	.2126	.2160